Who is the "Human" in Human-Centered Machine Learning: The Case of Predicting Mental Health from Social Media

STEVIE CHANCELLOR, Georgia Tech, USA
ERIC P.S BAUMER, Lehigh University, USA
MUNMUN DE CHOUDHURY, Georgia Tech, USA

"Human-centered machine learning" (HCML) combines human insights and domain expertise with data-driven predictions to answer societal questions. This area's inherent interdisciplinarity causes tensions in the obligations researchers have to the humans whose data they use. This paper studies how scientific papers represent human research subjects in HCML. Using mental health status prediction on social media as a case study, we conduct thematic discourse analysis on 55 papers to examine these representations. We identify five discourses that weave a complex narrative of who the human subject is in this research: Disorder/Patient, Social Media, Scientific, Data/Machine Learning, and Person. We show how these five discourses create paradoxical subject and object representations of the human, which may inadvertently risk dehumanization. We also discuss the tensions and impacts of interdisciplinary research; the risks of this work to scientific rigor, online communities, and mental health; and guidelines for stronger HCML research in this nascent area.

Additional Key Words and Phrases: human-centered machine learning; machine learning; social media; research ethics; mental health

ACM Reference Format:
Stevie Chancellor, Eric P.S Baumer, and Munmun De Choudhury. 2019. Who is the "Human" in Human-Centered Machine Learning: The Case of Predicting Mental Health from Social Media. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 147 (November 2019), 32 pages. https://doi.org/10.1145/3359249

1 INTRODUCTION

> The two hardest things in Computer Science are: People, and convincing others that "People" is the hardest thing in Computer Science. – attributed to Brad Grzesiak[1]

"Human-centered machine learning" (HCML)[2] is a rising subfield of computer science (CS) that combines the expertise of data-driven predictions and outside domain knowledge to make headway on questions of societal importance. These approaches have become popular in predicting elections [171], understanding criminal justice [175], and detecting fake news [163]; in HCI and CSCW, HCML has examined questions such as abusive content detection [40] and crisis [166].

HCML is focused on impacts to individuals, communities, and society, made explicit by its contributions to human-centered domains and challenges and self-stated goals within papers [25].

[1] https://twitter.com/listrophy/status/876129823130869760

[2] This emergent field has many names, including but not limited to, human-centered machine learning, human-centered AI, and data science for social good. For simplicity's sake, we use HCML as the umbrella term throughout this paper.

Authors' addresses: Stevie Chancellor, Georgia Tech, Atlanta, GA, USA, schancellor3@gatech.edu; Eric P.S Baumer, Lehigh University, Bethlehem, PA, USA, ericpsb@lehigh.edu; Munmun De Choudhury, Georgia Tech, Atlanta, GA, USA, munmund@gatech.edu.

2573-0142/2019/11-ART147 $15.00
https://doi.org/10.1145/3359249

However, this interest causes tensions in researcher responsibilities to the "human subjects" within their research [92, 123]. In machine learning (ML) and related areas, there are few protocols for managing researcher relationships to data [92, 187]; ML has historically relied on large and public benchmark datasets, like ImageNet [57]. Data now comes from sources much closer to the human – for instance, social media provides a large, unobtrusively collected source of data over time about peoples' thoughts, feelings, moods, and experiences. As data-driven research adopts human-centered research paradigms and moves closer to research traditionally protected through ethics boards [8, 72, 100, 105], this new proximity complicates the traditional representations of individuals involved in work from either ML or human subjects research perspectives alone. These *representations of the human* have downstream consequences on how research is conducted and reported, and how it may impact communities and individuals who are the object of study.

The impacts of these representations go beyond abstract notions of roles or responsibilities. Computationally focused work tends to treat individuals as data points to be analyzed, abstracting away from unique details to identify large-scale patterns and phenomena [113]. Predictions central to many HCML questions can both support decision-making and lead to important outcomes that pushes scientific understanding [63]. Yet, these same predictions also provoke extensive utopian and dystopian rhetoric [25, 112, 113]. Representations have meaningful consequences on research methods [13, 15, 170] and practical risks in increasing stigma [117], reproducing stereotypes and discriminatory practices [9, 134], and harming individuals and communities [59, 103]. As HCML emerges as a field with extensive academic and popular media attention, it is an opportune time to step back and assess its trajectory [98]. Explicit focus on these concerns can drive critical reflection in a nascent area to identify best practices [61, 92, 104] and, possibly, support redirection [17].

In this paper, we ask **"who is the human in human-centered machine learning?"** to explore these representations of human research participants in a new interdisciplinary space. We focus on language as operative in explicating these representations. Language is a driving force in how we conceptualize problems, include (or exclude) individuals from analysis, and encourage others to advocate for social change, as it dramatically impacts power dynamics and politics of oppression [74, 75]. The discursive representations of personhood can influence how people are justly and equitably treated [33, 99, 103, 129], and whether HCML research may fundamentally diminish respect and agency of those who are the object of study [85].

Work on predicting *mental health status* from social media data offers a prime area to study these representations. In this domain, computer scientists have designed algorithms to predict if someone is suffering from a mental disorder (*e.g.*, depression [54, 143]) or closely related symptomatology (*e.g.*, suicidal ideation [41]). This work is highly interdisciplinary across multiple fields of CS [28, 45, 140, 176, 178], predicting embodied illness from digital trace data. The topic is also sensitive and requires delicate care by researchers to not harm research participants [19, 36], providing an opportune setting to understand the representations of humans within HCML.

To conduct this investigation, we systematically identify 55 papers where CS researchers aimed to predict mental health status using social media data. We applied thematic discourse analysis to this corpus to examine how the human as data provider and beneficiary is described within these papers [75]. We identify five discourses that frame the human in varying, sometimes conflicting ways. Crucially, many of these framings result in a translation [34, 110], constructing the human as a data point for machine training and optimization rather than as a person who should be justly, equitably, and sensitively treated. A single paper will often invoke different discourses, leading to confusions and depersonalization. In short, the discourses within these papers weave a complex notion of who the human is and, in the process, inadvertently risk dehumanizing the individuals who are both the producers and beneficiaries of such analyses [85].

Our work highlights the tensions of representations of the human; we argue that these papers operate as "boundary objects" for HCML research [165], or documents that provide interdisciplinary flexibility between domains. Our findings surface a paradoxical representation within HCML of the human as being both in the "subject" and "object" positions, where humans are both centered and prioritized in the analyses but are also the object of machine learning techniques. We also put forward numerous scientific, community, and practical consequences for these representations, such as issues of reproducibility, the integrity of communities, and practical risks of dehumanization. We articulate these risks for dehumanization around Haslam's perspective, where actions "deny uniquely human attributes...and human nature to others" [85][p. 252]. Finally, we provide guidelines on how such representations can be made better, and we call for HCML researchers to adopt more human-centered practices within their work.

Reflexive Considerations. Two authors are social computing researchers who use ML to study mental health. Their research is included in the corpus, positioning them as both insiders to the subject area and object of critique. We value a human-centered approach in research that employs ML, and our perspective is informed by the primary commitment to improve societal outcomes. We are cautiously optimistic that HCML can assist in realizing these goals. Our ideas are guided by our experiences working with clinicians and psychologists and engaging with individuals who suffer from mental disorders. These experiences undoubtedly impact our perspectives on the critical analysis [21], and we view this position as "critical insiders" to be a unique and valuable opportunity to raise concerns and advocate for change.

2 RELATED WORK

First, we describe HCML and the contributions of multiple fields to its conceptualizations of the human. Then, we discuss our topic area – predicting mental health status in social media data).

2.1 An Overview of Human-Centered Machine Learning

Recent excitement and growth in HCML bely a longer scholarly history in AI, HCI, and solving problems at these intersections [162]. Grudin described the history of HCI and AI, the two fields alternating periods of flourishing while the other suffered a "winter" of reduced funding and researcher interest [80]. More than a decade ago, Winograd also outlined the strengths, limitations, and relevance of rationalistic and design approaches offered by AI/ML/data science and HCI when applied to "messy" human problems [180]. Indeed, Grudin's hypothesis about HCI and AI intertwining has come true nearly a decade later [80]. The focus of HCML echoes the desire to solve those "messy" human problems [180], and recent work in this domain for social media has been interested in complex areas such as abusive content detection [40], detecting fake news [163], and crisis informatics [166].

Complementing this rise in HCI-AI, scholars too have wrestled with the notion of human-centeredness and its connection to computing [8, 100, 105]. Human-centered paradigms for computing advocate for integrating "personal, social, and cultural aspects" [100] into the design of technology, and accounting for stakeholders in the creation of technological solutions. Yet, Kling and Star acknowledge that "there is no simple recipe for the design or use of human-centered computing" [105][p. 23], and Bannon acknowledges that different traditions of human-centeredness approach the commitment to human actors, as either straightforward interdisciplinary or as a deeper-seated commitment to a shared set of values [8].

We argue that HCML uses prediction systems in service of societal goals that focus on human needs and interests. However, because there is such interdisciplinarity under this umbrella and the field is very nascent, there are no formal definitions for what HCML is, nor shared vocabulary

within the field [27]. Different disciplines construct it in different ways – what comprises HCML and how does it represent the human actor in these disciplines?

In this section, we identify key developments that triangulate the definition of HCML. We focus on three contributions mapped to three domains: the *engineering* of algorithms to solve sociotechnical problems; the *responsibilities* of researchers drawn from research ethics and critical data science; and the focus on *interactions* of ML with humans from HCI.

2.1.1 Engineering for HCML. In addition to providing obvious methods foundations, ML has also placed its attention on engineering solutions to challenging social problems. The subarea of FAT (fairness, accountability, and transparency) makes contributions to HCML through algorithmic representations that promote better outcomes for values like justice, fairness, equity, and agency. In FAT, the focus is on engineered or computational solutions to these problems. This area has gained traction, as seen in the success of the FAT-ML workshop series (http://www.fatml.org/) and the recently conceived ACM FAT* (Fairness, Accountability, and Transparency, and others) conference (www.fatconference.org).

FAT approaches these problems as embedded within the data, methods, and algorithmic representations, and then uses statistical or mathematical techniques as a solution. For example, this involves identifying and controlling for undesirable biases and discrimination from ML and AI [65, 68]. In addition to questions about fairness and bias, the community is also interested in making algorithmic output more explainable [62, 106, 127, 144] and interpretable [89, 108, 126] by human actors involved in decision-making.

FAT's conceptualization of human-centeredness provides engineering solutions for algorithms that conform to these values like fairness, justice, and equity in prediction tasks. However, researchers within the area have cautioned against an over-reliance on these abstractions of sociotechnical problems; they argue mathematical solutions to abstracted notions of "fairness" may dramatically miss the social context of their applications [158]. We contribute to this conversation by focusing on this social gap, in our case the gap of representations of the humans who contribute data and benefit from HCML.

2.1.2 Researcher Responsibilities and Roles. There have also been broader calls to consider the social, cultural, and ethical responsibilities to humans within data-driven research paradigms [25, 123], in a new area termed "critical data studies." For HCML, we focus on critique around the scientific representations of the human and engagement with "participants" in big data research.

Historically, humans in research were human subjects with rights guaranteed by an ethics board that managed appropriate scientific conduct. For the United States, this manifests in the Belmont Report wherein the research must align with three values: beneficence, justice, and autonomy [72]. In contrast, ML has focused on large, publicly available datasets, like the famous ImageNet [57], and has not directly engaged with humans.

In HCML and other data-driven areas, the roles and relationships of scientists to research subjects has become hazy, as datasets are now curated for the relevant domain. On social media predictions of human behavior, where the data and prediction target are both the object of interest, these tensions become more pronounced [92]. We see scholarly debate on this subject, which has challenged conceptualizing who the human research subject is in these scenarios [123, 187]. Numerous researchers have questioned meaningful protocols of gaining consent [97, 187, 188], and the places where researchers agree and disagree around protocols with social media data [174]. Considering the other side of these relationships, individuals' perspectives on whether Twitter research is ethical is also hazy. Empirical research has shown conflicting and highly contextual opinions of individuals to have their Twitter accounts used for research [71, 124]. Public reactions to data privacy violations by these social networks are also mixed [70, 137].

We build on this approach to human-centeredness, which considers the roles and relationships between scientists and the individuals involved as key to understanding the representations of the human in HCML.

2.1.3 Human Interactions with Technology. HCI has also been interested in the use of ML to augment interactive and intelligent systems [2, 91, 183], emblematically seen in interest in automated chatbots [179] and criticisms of these practices [154]. Critical mass around HCML is also forming in HCI, measured by interest in workshops at recent CHI and CSCW conferences [3, 76, 98].

Of most interest to us, HCI's and CSCW's intellectual foundations have also preoccupied themselves with meaningful representations of humans, people, organizations, and communities and their interactions with technical artifacts. HCI has historically delegated these relationships and interactions to the notion of the "user"[15, 44, 109]. These representations have been challenged at various times through HCI's history to better represent who the user is (and is not) [15, 151], how to contextualize social roles for users [109], and what user identities appear in this research [153].

Our work speaks to larger questions of representation and research interest within the CSCW community. CSCW and social computing have a history of considering the social, organizational, and cultural contexts in which humans act [7, 58]. We situate our work between numerous contingencies within CSCW: a history of human-centered machine learning work [39], critical perspectives on data science [103], and exploring the ethics and representations behind this work [174]. We build on prior work in providing the first case study of how the human is framed by the scientific publications that describe HCML.

2.2 Predicting Mental Health Using Social Media Data

Next, we discuss the research area in our study, predicting mental health status using social media data, and our motivations for selecting this area.

Since 2013, computer scientists have designed algorithms that can predict with high accuracy if someone is suffering from a mental disorder or related symptomatology using social media data (for a thorough overview of the field, see Chancellor *et al.* [36]). These algorithms predict if someone is suffering from a mental disorder, like depression [54, 140, 177], anxiety [160], post-traumatic stress disorder [47], schizophrenia [23, 125], and eating disorders [50, 176]. These approaches are also sensitive to symptomatology related to mental disorders, like suicidal ideation [41, 90, 121, 135], stress [116, 185], and the severity of mental illness [37, 156]. Research heavily adopts data-driven prediction methods from ML, like supervised binary classification [31, 94, 101], regression analysis [156], and recently deep learning [20, 77]. Taking inspiration from [36], we refer to this area as *predicting mental health status* to encompass both predictions on disorders and their closely related symptomatology.

We chose to focus on this domain for several reasons. This work has received increased attention [157, 182], not only within HCI and CSCW [37, 55, 168], but also in multiple subfields of CS, such as natural language processing (NLP), machine learning, and medical informatics [28, 45, 140, 176, 178]. This attention is also reflected in new literature reviews and meta-reviews, which examine practices within HCI for affective disorders research [149], for qualitative research [157], and for data-oriented approaches [182]. Given its focus on health and machine learning, there is also a growing history of critique and concern for using online trace data to predict health outcomes [19, 43, 124]. The implications of this research posit numerous social benefits – in new monitoring and public health efforts, designing interventions for dangerous behaviors, and potential to improve health and well-being. Finally, this research focuses on predicting intrinsic, bodily characteristics on the individual – mental disorders are manifested in people's physical well-being,

Table 1. Keywords for literature search.

Category	Keywords
Mental health (1)	mental health, mental disorder, mental wellness, suicide, psychosis, stress depression, anxiety, obsessive compulsive disorder (OCD), post traumatic stress disorder (PTSD), bipolar disorder, eating disorder, anorexia, bulimia, schizophrenia, borderline personality disorder (BPD)
Social media (2)	social media, social network, social networking site, sns, facebook, twitter instagram, forum
Search term	**(1) AND (2)**

and thus data science efforts to understand them are inherently humanistic. Thus, it offers an opportune setting to consider how different CS fields construct a representation of the human.

3 METHOD

We adopted the tools of a systematic literature review to gather studies, informed by general standards for literature/meta reviews [114] as well as those in HCI and CSCW [12, 60, 61].

Constructing a corpus across disciplinary boundaries in CS is difficult. We could not use a single professional organization's search database (ACM or IEEE); however, most scholarly indexing services, like Web of Science or Scopus, do not consistently index CS conference proceedings. When we tested our initial search strategy through these services, journal entries were robustly indexed; yet there were large gaps in the coverage of conference proceedings that are important in these areas (*e.g.*, AAAI, ACL, CHI, NIPS/NeurIPS, DH, AMIA). Initial experiments with keyword searches through engines like Google Scholar yielded an intractable number of results (over 200,000 candidate papers before deduplication).

For our approach, we iteratively generated a list of 41 venues (both conferences and journals) that "seeded" our search. We used keywords to filter in these venues, then identified candidate papers through this list. Finally, we sampled the references of these papers once to identify missing papers from the first pass. This produced 55 papers in total. Our methods are summarized below[3].

3.1 Searching the Datasets

First, we searched the literature in May 2018 for articles published between 2008 and 2017, dovetailing with the emergence of academic research on social media [26].

We developed two sets of keywords to search in pair-wise fashion: those for mental health and those for social media. These were inspired from meta-reviews on social media and mental health [157, 182] and our expertise in the area. A list of our keywords can be found in Table 1.

Next, we searched for these keywords across 41 English venues in the interdisciplinary intersection of prediction of mental health through social media. These were inspired, again, by our expertise in the field as well as from the results of previous literature reviews in the space [157, 182]. A full list of venues is in the Appendix (section A.1).

Three search engines were used to ensure robust coverage across venues. We used the ACM Digital Library for ACM journals and conferences, Google Scholar using the Publish or Perish software [84] for other conference publications, and Web of Science for journals[4]. One venue (CLPsych) was not indexed correctly by any search engine, so we manually searched the proceedings

[3]More details on our methods can be found in the Appendix
[4]www.webofknowledge.com

for matching keywords in the title and abstract. We identified 4,420 manuscripts that matched these keyword pairs.

3.2 Filtering Strategy

We first filtered the manuscripts to include peer-reviewed, full-scale archival studies published between 2008 and 2017, deduplicating entries as we went. We honored the home community standards to assess archival status[5], including studies that conduct full-scale research as primary sources. This removed meta reviews and literature reviews, news reports, case studies, panel proposals, and shared tasks. After deduplication and filtering, this produced 2344 manuscripts.

Next, we manually filtered by title and abstract, removing spurious items obviously not related to mental health or social media data. Examples of mismatches included other health conditions, such as cancer or diabetes, or data sources such as electronic health records. This reduced our corpus from 2344 to 87 papers. Finally, we read and fully screened all 87 papers, using the following criteria for inclusion in our analysis of HCML in this domain:

(1) They must address mental health in clinically specific ways. This meant studying a mood or psychosocial disorder (*e.g.*, depression, anxiety, schizophrenia), symptomatology from the DSM-V [5] about disorders (*e.g.*, suicidality, psychosis), or the severity of mental disorders (*e.g.*, moderate vs. severe depression). We excluded subjective mood, well-being, happiness, or general emotions not directly related to mental disorder diagnosis. We also excluded papers about mental disorders and conditions that are not mood or psychosocially-oriented (*e.g.*, ADHD, autism spectrum disorder) [5].
(2) The paper's method must focus on quantitative prediction through ML techniques from social media data. This included regression analysis, machine learning, and time series analysis.
(3) The paper must study social media data from social networking sites, blogs, or forums. We excluded other digital data traces, such as search engines or app use (if not related to social media apps).
(4) Finally, the prediction must be made on an individual. We excluded papers that made aggregated predictions on groups or communities to effectively scope the literature review for precision around the same kinds of research questions and contributions, aimed at directed care and interventions for individuals. If a paper made predictions on individuals later aggregated for another purpose, we included these.

This process generated 44 papers that matched all of our constraints. Finally, we conducted an iterative pass, sampling related papers to our 44 identified from the bibliographic details of the citations. We then undertook the same screening and filtering process above, moving from 519 candidate papers to 11 papers that matched our constraints. Additional snowballs through these papers did not return substantially new results.

This produced 55 papers (44 from initial analysis + 11 from iteration) included in this analysis. The full list of all 55 papers is provided in the Appendix. We will call this our *corpus* throughout the remainder of the paper; individual papers from the corpus will be described as a *document*.

4 ANALYSIS TECHNIQUE - DISCOURSE ANALYSIS

To understand how our case of HCML conceptualizes responsibilities to humans, we study how the community describes them in publications, or the *discourse*. Foucault famously described discourse as an action of language "that systematically form[s] the objects of which they speak" [74]. Discourse frames, shapes, and changes our formations of social and political structures, and how power and

[5]In CHI, workshop proceedings are not considered archived; however, in ACL, workshop proceedings that appear inside the ACL Anthology are archived.

responsibility may be conferred to individuals and groups – with the ultimate goal of making such structures apparent for critique and change [74, 75, 87, 102].

Discourse has been a useful lens to understand language focused on the adoption and use of technology. In HCI, focuses on language and representation have been used to explore lay narratives around robots [167] and smartphones [83]. Hoffmann has explored the pitfalls of anti-discrimination and anti-bias discussions in data science research and practice [87], and Hoffmann and collaborators explored how Facebook's CEO, Mark Zuckerberg, changed his conceptualization of the relationship between Facebook and its users during his tenure [88]. Discourse has been a useful frame for critically considering practices within HCI itself, such as intersectional identities of research participants [153] and the field's construction of sexuality [102].

Driven by our primary research question (who is the human in HCML), we identified research sub-questions to better examine this. These included:

- Who is the human or subject of these predictions represented in the paper?
- How are these subject positions represented? [15]
- Are there notable proxies or substitutes for the notion of the human in the corpus?
- Who are the benefits/implications of this research offered to?

Using inductive coding [75], the lead author conducted a close reading of the entire corpus, annotating the terms and phrases that conceptualized the human "research subject." She focused on the subject of the prediction task, the studies, as well as the purported beneficiaries. She coded at the sentence level for terms and phrases, as disambiguating at word-level was too granular to draw larger conclusions. Thus, statements could be simultaneously coded for the presence of multiple terms or concepts. The author also wrote notes/memos from insights gleaned from close readings and thematic corpus-wide observations. As the coding progressed, the authors decided to not code Literature Reviews/Related Work, as these sections reported on other studies' representations.[6] This portion of the analysis was done in Dedoose[7]. After the initial coding was complete, the authors then met and discussed emergent themes [29], which we identified as discourses governing the representations and relationships of the human within the corpus. These observations and understandings form the basis of our analysis presented below.

5 FINDINGS

5.1 Discursive Representations of the Human

We identified 164 novel terms that describe the human across the corpus. We then grouped terms based on the ways they were used in the documents, which produced five discourses: Disorder/Patient, Social Media, Scientific, Data/Machine Learning, and Person. We clustered terms used in more than one document, as we felt this was more emblematic of patterns in the corpus. We present an overview of these discourses in Table 2. In the following sections, we unpack the representations of who the human is within HCML in these five discourses.

5.1.1 Human as Patient/Disorder. To begin, we found many conceptualizations of the human as a clinical subject, emphasizing their relationships with disorders, doctors, or clinical researchers. This category had the most variety of terms in our dataset.

One of the most common patterns of language use was referring to the human as a "patient." Using measures such as self-reported clinical status, researchers referred to individuals as if they were in an active clinical care relationship. For instance, Nakamura *et al*. analyzed 200 authors of

[6] We believe the transmission of these roles throughout literature reviews is ripe for future work, but was outside of the scope of the present study.
[7] https://www.dedoose.com/

Table 2. High-level discursive categories from our analysis oredered in decreasing order by appearance in unique documents. We excluded words used in only one paper.

Discourse	*Terms (number of documents/papers)*
Disorder/Patient	patient (17), depression (10), depressed (9), sufferer (9), behavior (7), condition (4), distressed (4), PTSD (4), neurotypicals (3), non-depressed (3), suicide (3), normal (3), victim (3), clinical (2), anxiety (2), bipolar (2), mentally ill (2), non-stressed (2), pro-anorexic (2), stressed (2), suicidal ideation (2), score (2), standard (2), state (2)
Social Media	user (55), post (25), tweets (16), content (15), account (14), author (14), community (10), microblog (7), text (7), document (6), member (6), activity (4), followers (4), message (3), poster (3), tweeter (3), corpus (3), blog (2), item (2), networks (2), publisher (2), profiles (2), lexicon (2)
Scientific	population (29), control (21), participant (16), subject (10), cohort (8), candidate (6), respondents (6), observation (2), pool (2)
Data/Machine Learning	data (31), sample (25), dataset (18), class (16), example (8), subset (8), test set (5), category (4), positive/negative (3), task (3), data point (2), model (2), prediction (2)
Person	people/person (47), individual (40), she/he (11), woman (7), one's (5), man (5), youth (5), student (5), mother (4), worker (4), crowdworker (4), female (3), someone (3), peers (3), friends (2), others (2), they (2), adolescents (2)
Not Grouped	group (35), case (9), counterpart (2), life/lives (2)

blogs tagged depression from a Japanese health blog portal. Throughout the paper, they refer to individuals who write the blogs as "long-term patients" [128] — the title of the paper, "Defining patients with depressive disorder by using textual information" reflects this decision.

The term "patient" may not accurately reflect the relationships these individuals have with clinical care providers. A patient is someone who is actively participating in a health care relationship — no studies in our corpus actively recruited participants through clinical practices or formally verify a relationship with a health care professional around a mental disorder. A few studies verified clinical diagnosis or date of treatment [51, 54], though these studies did not call individuals "patients."

We also noticed diverse language describing disorder status and the individuals grouped under it. One common pattern was to use the language of the disorder as shorthand for the positively identified group. For example, authors asserted that individuals identified through proxy measures actually suffer from that mental disorder, and then use that language in the remainder of the paper. In one document, Shen and Rudzicz used participation in r/Anxiety (a subreddit for anxiety in general) as a signal to identify the "Anxiety group":

> "we also find lexicons relating to feelings and first person pronouns...represented in the *Anxiety* group" [160][p. 63]

Crucially, anxiety is an overloaded term; it can mean an emotional state or short-term experience, a symptom of other disorders, or the category of anxiety disorders. Therefore, defining participants as the "Anxiety group" may be misguided for clinical purposes.

We identified similar patterns in individuals with presumed absence of the mental disorder of interest. Many papers adopted the language of "non-disordered," to contrast a group of "disordered" individuals, such as the "non-stressed user." [115]

We also saw discourse framing the individual who did not have a mental disorder as "normal":

> "We perform an empirical study...of potentially depressed users against a differential control group of *normal users*." [173][p. 135]

or "neurotypical":

> "Users who attempt to take their life generate tweets at a level higher than *neurotypicals*" [48][p. 113]

Using language like "normalcy" or "neurotypicality" to describe a lack of a mental disorder stigmatizes those who have mental disorders by othering them and their experiences [49, 117]. Several guidelines written by both journalists and mental health advocacy groups suggest avoiding language that paints the individual as just a mental disorder [49, 117].

Overall, this discourse cast the human as active participants within clinical relationships, implying engagement with clinical partners ("patients," "the depressed") or with language that can be stigmatizing for individuals who suffer from mental disorders.

5.1.2 Human as Social Media. The second discourse we found was social media as the mediating actor in these relationships. We begin by looking at the term "user," present in *all documents* in our dataset. It most commonly referred to the "user" in relationship to a social media platform like Twitter: "We extract several features from the activity histories of *Twitter users*." [168].

We saw similar patterns around the more generic term "poster" – "this distribution was skewed by a smaller number of frequent *posters*" [141], or platform-specific language like "tweeter":

> "this paper proposes to leverage details of social interactions between *tweeters* and their following friends" [185][p. 26]

In these contexts, the human was portrayed as an active curator of their social media profiles who generated data or interacted with others on the platform.

In contrast, we also saw representations framing more passive engagement. Many documents described the entire collection of social media data as the object of prediction, or the individual units of engagement, using language like "account," "profiles," or "posts":

> "All potential control Twitter *accounts* were also manually curated." [122][p. 123]

Many documents used a single positive identification of mental health status on these passive sources of "post" or the "Tweet," then scaled it to the human behind the account. In one example, the authors described how they can detect mental disorders in people and in populations, though they only use a single post in several mental illness Reddit communities to draw that conclusion about an individual [77]. A single episode of a behavior or symptom measured through a "post" may not be not enough information to comprehensively identify mental health status. This post-to-human proxy transformation was subtly implied throughout the writing, thought rarely explicitly stated.

We saw this pattern consistently throughout the documents – Homan *et al.* used Twitter "microblog text" to predict suicide risk, identifying moments for urgent intervention [90]:

> "distress is an important risk factor in suicide, one that is observable from *microblog text*" [90][p. 108]

However, none of the papers in the corpus substantiated how a single moment of distress (as measured by a single post) may communicate urgent risk or the presence of a larger disorder. This is worrisome, given that the vast majority of documents in the corpus that reduce their observations to single posts rarely contextualized outside of a single unit of observation about an individual.

Finally, we also noticed interesting contextual compression around online communities. Often, participants in an online community are assumed to have the mental health status of interest. In this

example, Masuda *et al.* used the proxy signal of topical community membership as an indication that someone is suicidal:

> "The dependent variable that represents the level of suicide ideation is binary, i.e., whether a user belongs to a *suicidal community* or not." [121][p. 6]

We find that there is a logical assumption that participation in a community is indicative of mental health status. However, communities are nuanced venues for participation – individuals have varying needs and reasons why they engage with communities, some of which may not indicate that they actually are afflicted by the condition of interest. Often in our corpus, the term "community" was a labeling tool or mechanism for the disorder of interest, a problematic representation of mental health status [67].

To summarize, we found the discourse around social media use to both promote active and passive engagement of the humans. Social media is one insight into well-being and cannot comprehensively represent individuals' thoughts and behaviors; thus, examining it requires a necessary compression of fidelity. However, we found that many papers overcompressed the representation of mental health status of individuals or communities to a single behavior, message, or post on social media.

5.1.3 Human as Scientific Subject. We move to the third discourse, drawing on perspectives of the human as a scientific subject. To begin, one popular representation was the human as "participant."

In some studies, individuals provided researcher access to their social media [136, 148]. For instance, Guan *et al.* recruited over 900 participants through recruitment messages on Sina Weibo: "All *participants* interested in this survey were asked to log on to the Internet survey system by their Sina Weibo account." [81][p. 2]

Scientific language for "participation" has evolved around active consent into research through ethics boards protections [72]. However, "participant" and the closely related "subject" were not always used precisely to refer to human research subjects; in fact, several studies used this language to denote individuals passively gathered from public social media data:

> "...[This] evaluation demonstrates that our system can effectively identify potential *subjects* who are suffering depression but are unaware of it..." [161][p. 285]

We saw similar confusion with scientific terms like "control," referring to a group of individuals juxtaposed against the positively identified group.

> "Data...distinguish[es] users with schizophrenia from *healthy controls*" [23][p. 1]

However, "control" was always juxtaposed against an implied "treatment" position — in experimental setups, control groups are verified to not have the effect under observation. In many documents, authors would draw a random sample of users from the rest of the site and use this as their "control." Although some studies use screeners with consenting participants to evaluate this [28, 54, 184], most did not validate their control group.

This mathematically guarantees that the "control" is not a true control group, as it will possess those who have the mental disorder of interest, given the occurrence rates of mental disorder in the general population. Coppersmith *et al.* reflect on the problems of contamination and explain their language choice for "control," saying, "...we draw an age- and gender-matched *control* from a large pool of random English users." [48][p. 109]

We saw similar dilution of terms like "population," which can have scientific meaning as well as statistical relevance. Often, documents referred to "population" as the whole group of individuals involved in a study, as Saha *et al.* do for studying all content from a campus community: "we proposed computational techniques to assess how the psychological stress of a campus *population* changes following an incident of gun violence." [148][p. 24]

However, language around "population" can obfuscate whether the authors had the whole universe of people of interest to the research (*e.g.*, all people who participate in a depression community) or a sub-population of that group (*e.g.*, 500 people sampled from a depression community). Chancellor *et al.* make this confusion with "user population" to refer to a sub-population of users to evaluate trajectories of anorexia recovery, "after 45.6 months, 50% of the *user population* have not recovered." [38][p. 2116]

In sum, we found that scientific discourse was incorporated into these studies in imprecise ways. Many studies will borrow terms from the experimental or human subjects literature ("participant," "subject," "control"), implying experimental rigor and human subjects protections that are not realized through the actual experimental design.

5.1.4 Human as Data/Machine Learning Object. The fourth discourse we identified is the human as data or ML object, translating the person into a part of an algorithm or machine learning pipeline.

We begin with the word "example." Here, Benton *et al.* used "example" to references the mathematical number of data points passed to their algorithm as a key component of their contribution, "we show how to model tasks with a large number of positive *examples* to improve the prediction accuracy of tasks with a small number of positive examples." [20][p. 153]

Other data terms defined the transformation of human to data object, such as "positive"/"negative":

> "*Positive*: the tweet content indicates the presence of one of the studied diseases/states in the person who has written the tweet." [139][p. 5]

When describing the methods or results, mathematically precise language can carefully identify what data is incorporated into machine learning algorithms. Some studies are aware of this distinction and drew attention to it: "Because we did not interact with our subjects and *the data* is public, we did not seek institutional review board approval." [37][p. 1172], but most do not.

However, data discourse was often used ambiguously to describe the contributions of the paper without describing how the data was ascribed to a person. Some examples referred to changes in groups of people, or "classes," despite predicting on individuals:

> "for the depression *class*, we observe considerable decrease in user engagement measures" [54][p. 133]

These contextualizations can be useful for understanding group behavior, but must be kept in context to the operative research question of predicting on individuals.

In another example, the authors provided research contributions without referring to individuals:

> "data mining of online blogs has the potential to detect meaningful *data* for depression studies. The result highlights the potential applicability of machine learning to psychiatric practice and research." [131][p. 224]

Finally, the situating language in the sentence highlighted this division of the human to the data:

> "our work aims to make timely depression detection via harvesting social media *data*" [159][p. 3838]

Rather than collecting or gathering data, the researchers described their process as "harvesting."

Extreme care must be taken around proxy language that converts individuals' social media content into data for machine learning observations, as it risks objectifying dehumanizing individuals in these documents. These discursive choices can make invisible individual experiences or imply that the research is not actually interested in serving the individuals it argues it helps.

5.1.5 Human as Person. The final category of discourse was focused on characteristics and roles of people. One common term from this category of terms was "individual," framing the contributions: "A technique to identify symptoms of depression in *individuals* from objective information would

hasten recognition of depression" [168][p. 3187-8]. We also occasionally saw the use of "individual" in the prediction analysis: "using the same feature set can build a classifier to classify depressed *individuals*" [173][p. 128]

We observed similar translational challenges to the social media language with the post-to-individual proxy, where presence of behaviors in posts are implied to affect the individual. In this document, the authors built a decision tree to predict suicidal ideation on posts, implied that it transfers to the individual:

> "The tree first splits on the "achieve" category of LIWC, such that if an *individual's* usage rate of achievement-related words exceeds 1.46, that *individual* is labeled as nonsuicidal." [28][p. 5]

Other documents made similar proxy assumptions, arguing that they correctly identified "people" who are depressed without clinically verified proxy signals and explaining the findings in light of that: "depressed *people* sometimes suffer occupational function impairment, which leads to different mental conditions or behaviors between workday and weekend." [161][p. 280]

We also saw reference to the roles and demographics of the individuals of interest in the study. Several studies examined the social media behaviors of "students," "teenagers," or "adolescents," and used these terms throughout:

> "36 *high school students* (15 males and 21 females, aged between 15 and 17) in Shaanxi Province, China, participated in the user study." [185][p. 32]

Another demographic focus was on gender. One instance we saw was use of the term "mother" to refer exclusively to birth mothers who have postpartum mood changes:

> "the total timespan of our dataset is between March 2011 and July 2012, with a total of 36,948 posts from the 376 *mothers* during the prenatal period" [51][p. 3269]

However, gendered terms were also used imprecisely ("she," "her"). We saw this, especially in discussions of eating disorders, *e.g.*, "If *she* does not recover in 3 years, the probability of remaining anorexic for another 3 years is 0.39/0.56 = 69.6%" [38][p. 2116]. In this case, the term "she" was used throughout the paper and "he/him" was never used. Either the authors were using she/her in place of a pan-gender pronoun, or they were implying that the dataset was composed of women.

In another document, the author collected a dataset of "females" who do not self-report to have eating disorders as the negative dataset:

> "As ED develop predominantly in young *females*, the effects of demographics can be further controlled in comparing ED and Younger user" [176][p. 94]

This choice is concerning because the use of gendered or other demographic language coupled with mental disorders can reproduce stereotypes about who has certain conditions. About a third of those with eating disorders are male[8], and describing eating disorder sufferers as exclusively female is alienating to men with eating disorders. This could be applied to other mental health statuses that may perpetuate stereotypes.

In summary, we believe that the use of person-centered discourse in this paper is complex and may not always be directly tied to the person, involving proxy transformations of data to the person involved or overgeneralizing who suffers from a specific mental health status.

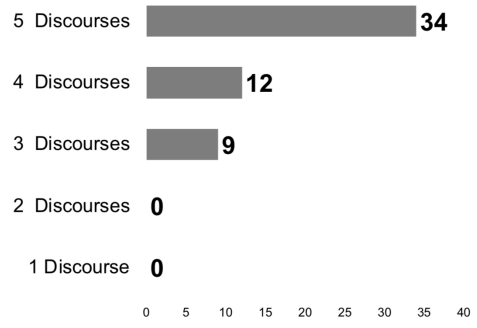
5.2 Relationships Between Discourses

Previously, we examined each discourse as an independent unit of analysis. Now we explore the interactions of these patterns within our dataset, as the interplay of these themes reveal larger trends in the way the representation of the human is constructed.

[8]https://www.nationaleatingdisorders.org/learn/general-information/research-on-males

In Figure 1, we show the number of discourses present in papers. All papers have at least three discourses present, and the majority (46/55) have four or five present in the writing. This indicates that these discourses are interacting in the majority of the documents.

Fig. 1. Counts of Discourses Present in Documents



To explore these relationships, we identified documents that had high discursive coherence, where one or two discourses were dominant and others were used sparingly. Very few papers had strong discursive coherence.

One document with high discursive coherence was De Choudhury *et al.* [51], who investigated extreme emotional and behavioral changes to indicate risk for postpartum depression. The preferred term for the human was "mother," even in the data analysis, results, and methods:

> "...for the volume measure, *mothers* in the extreme change class (C1), exhibit median change of -0.88 postpartum, indicating an 88% drop in posts per day..." [51][p. 3271]

Another example was Reece *et al.* who consistently referred to humans with Scientific and Person-Centered discourse [141]. They used "participants," "individuals," or "observations" throughout:

> "For the depression study, we analyzed 74,990 daily observations (23,541 depressed) from 204 *individuals* (105 *depressed*)." [141][p. 5]

Jamil *et al.* also had strongly consistent discourse patterns throughout the paper, though they used the terms "user" and "at-risk" to refer to humans in the dataset at risk of depression:

> "we trained a user-level classifier that can detect *at-risk users* that achieves a reasonable precision and recall." [101][p. 32]

Despite the variety in what discourses were the most prominent, we noticed that papers with high discursive coherence had increased attention to the individuals involved. This was signaled in the documents by details of data collection and ethics board approval (or lack thereof), ethics and privacy considerations, more detailed Introductions and Discussions, and potential negative consequences of this work.

However, for the majority of the corpus (46/55), most documents use four or five of the five discourses throughout the document. In one example, the authors moved between Clinical, Social Media, Data, and Scientific language in a single sentence, describing the impacts of their work:

> "While we used all *users* posting on mental health subreddits, only a *subset* of *authors* appears in the *control dataset* (around 9% of the *users*; 32,280 appear in the *non mental health* subreddits and 348,040 appear in the mental health subreddits)" [77][p. 6]

Other documents displayed increasingly complex and confusing representations of the human at the sentence level: "Data from disclosures deemed true were used to build a classifier aiming to distinguish *users with schizophrenia* from *healthy controls*." [23][p. 1]

Here, the individuals identified to have schizophrenia were represented as the generic "users," though those that do not were considered "healthy controls"; the reason for these distinctions of the individuals is unclear. These differences make it difficult to follow the sources of data within the paper, and could imply a difference in the sampling strategies for different groups.

More commonly, however, was a distinction between the framing of the individuals involved in the data process and the intended beneficiaries of the research. These documents framed beneficiaries in the Introduction and Discussion sections as "sufferers" or "individuals," indicating broader societal impacts with more humanizing language.

As an extended example of these patterns, the authors of [176] described how their approach may identify individuals who suffer from eating disorders and improve monitoring and detection. In the Introduction section, they use Person-centric language:

> "We sample *individuals* who self-identify as ED-ed in their profile descriptions on Twitter...thereby providing guidance to develop effective interventions not only for *individuals* but for large groups." [176] (Introduction)[p. 92]

However, the paper shifted to describing the data, methods, and results oriented around the most dominantly used term, "users" and data-oriented language ("sample," "dataset"):

> "We first present three types of measures to characterize differences between ED-ed and non-ED-ed *users* on Twitter" [176] (User Characterization)[p. 94]

In these examples, we find there is a malleability of roles when many terms are used to describe humans, as their roles changed based on the implied needs of the researcher. These shifts in documents with low discursive coherence subtly constructed the relationship of the researcher to the individuals, treating the individual as a flexible component of the research agenda.

Our findings reveal the importance of discursive framing for building the representations of the human within HCML. Using techniques from discourse analysis, we identified five discourses in our corpus: Patient/Disorder, Social Media, Scientific, Data/ML, and Person, as well as novel interactions between these discourses in constructing this representation. These discourses reveal numerous gaps, inconsistencies, and potential harms that we explain in the next section.

6 DISCUSSION

From the 164 distinct terms in our analysis, five dominant discourses appeared. The benefits that this research can provide for humans is articulated in these papers' Introductions and Discussions, using Person-centered discourse around "people" and "individuals." However, the technical portions of many papers refer to these humans as data objects ("sample," "positive"/"negative"), as well as with language that confers diagnostic authority within these models by calling individuals "patients." Terms and meanings frequently meshed together, even at the sentence level, and different discourses competed throughout the documents.

Our research surfaces a unique paradox within HCML of placing the individuals and communities involved in multiple, paradoxical positions that risk dehumanization. Next, we explain these findings in light of theory and how our case study of HCML provides unique insights into consequences of a human-centered research paradigm.

6.1 Contextualizing the Findings of this Work

In our case study of HCML, the documents function at several interdisciplinary intersections: health, machine learning/prediction, social computing, and data science. We suggest that these documents likely act as "boundary objects" [82, 165], negotiating the tensions among the "social worlds" of different disciplines that participate in this research. In each of these disciplines, the same words – *e.g.*, data, model, sample – may have different meanings. Star and Griesemer argue that actors must "reconcile these meanings if they wish to cooperate" [165][p. 388]. We interpret the shifting, varied, and inconsistent discourses as evidence of attempts at reconciliation. As a virtue of being a boundary object, these negotiations result in translations [34, 110], or explaining their understandings in one field to others and outsiders. These translations often represent the

necessary ontological abstractions needed to transform nuanced behavior into other domains [34], such as rigid computational structures represented by databases, regression models, and neural networks [30]. Given the areas involved, it is not surprising that the discourses construct the human according to practices within these disciplines.

However, translations come with risks. Here, a prominent risk is inadvertently *dehumanizing individuals.* Haslam describes dehumanization as, among other things, "an abstract and deindividuated view of others that indicates an implicit horizontal separation from self, and a tendency to explain the other's behavior in nonintentional, causal terms" [85][p. 262]. In our case study of HCML, the discourses used in the literature construct the subjective, complex human subject related to mental health as a data point for machine analysis, modeling, and diagnosis. At the extreme, we noted the use of terms such as "harvesting," "exploiting," or "extracting" information from a human, aligning with the kind of abstract, deindividuated view that Haslam describes. Similarly, the kinds of mathematical and computational techniques used for prediction often arrive at "nonintentional, causal" explanations for an individual's mental (health) state.

Related areas to HCML have identified potential consequences of such abstractions and simplifications. In FAT, for instance, researchers have warned of the overemphasis of abstractions of social complexity for mathematical aims [158]. In HCI, scholars argue that simplifications and proxy language have emerged that may elide certain complexities of the "user" experience [13, 15, 44, 109, 151]. And in health more broadly, such concerns have emerged as a shift from disease-centered medicine that diminishes patient agency to advocating for patient-centered care [18, 66].

From our case study, we argue that HCML uniquely risks dehumanizing individuals because of the paradoxical contrast of its human-centered commitments and the ways of knowing in AI and machine learning [32]. Human-centeredness explicitly calls for those who design technology to put the human at the center of their concerns [8, 105]. This recentering invokes the Foucauldian "subject" who has agency and power within relationships between actors [15, 74]. However, machine learning and the areas it draws on (*i.e.*, statistics, computer science, optimization research) view "the topic of study as an object" [27][p. 375]. The mathematical and computational techniques used for prediction often arrive at detached and causal explanations for an individual's mental health state. Our findings surface this distinction because the topic of study is both the individual (as subject) *and* the abstracted model/data point (as object), fundamentally at epistemological odds with each other. By "objectifying" the human for translations, this detachment separates researcher from the self, the inherently human focus identified by Haslam, and therefore risks dehumanization [85].

6.2 What Are the Consequences?

Our analysis reveals that this body of research risks dehumanizing the humans involved in HCML. For mental health, this tension surfaces in part because the person and the mental disorder – literally tied to their physical body – is the self-stated interest and empathetic goal of the researchers. To be clear, we do not believe the researchers conducting this work are intentionally dehumanizing the humans involved in this research. It is apparent from the humanistic and Person-centered discourse mentioned in these papers' Introductions that they are motivated by benefiting patients and people. We suggest that, in trying to perform the difficult translational work necessary to conduct this particular kind of interdisciplinary research (*i.e.*, HCML), the discourses that emerge in these papers end up inadvertently dehumanizing an already vulnerable group the work was intended to help.

Inconsistent representations of the human have practical consequences for both the research and the individuals involved. Just as discourse constructs notions of agency and power [74, 75] and influence lay opinions with downstream impacts on policy [132], these representations imply responsibilities, ethical decisions, privacy protections, and other obligations researchers have to

the humans represented within their datasets. Below, we discuss potential risks and consequences emerging from these inconsistent and sometimes inaccurate descriptions.

6.2.1 Scientific and Collaborative Consequences. Disciplinary knowledge manifests in narrative framings and specialized language through dissemination processes, like publications [10, 11]. Authors have an incentive to match the styles and practices in a venue to get published. We argue that venue fragmentation within this area may lead to inconsistent scientific and collaborative standards – 30 venues were represented in our final dataset across a wide variety of subfields (*e.g.*, HCI, health informatics, NLP, and AI). Reviewers may not know how to navigate different topic areas; there are few reviewers who are experts in mental health, social media, and machine learning at the same time, and also can review for all venues of interest.

What are some outcomes of these papers operating as boundary objects? First, these practices jeopardize reproducibility, a core value of much empirical research. Describing the representations of the human with shifting and poorly explicated terms can make it challenging to understand key questions for study design, such as inclusions and omissions of data. During our close readings, we struggled to understand the source of data based on shifting language in some documents. This relates to recent concerns around establishing construct validity of "proxy" signals to measure mental health [67]. Are we, through our reporting practices, ensuring that future work can appropriately measure the phenomenon of interest?

When discourses compete in these documents, framings may not explicitly establish when to adopt certain concepts. Take the example of patient: doctors have expectations for a "patient," who is an individual receiving treatment and under the care of a trained health professional. When used in a medical science paper, "patient" serves as a boundary maintenance mechanism [172]. However, similar translations of "patient" into HCML (when the concerned individuals do not meet the medical definition of a patient) can make the work difficult for outsiders to understand. Furthermore, these translations will likely be incorrect for downstream application into real-world technologies and treatment paradigms. These shifting concepts impact interdisciplinary collaboration, a critical component to mental health prediction and HCML in general.

Finally, complications can arise when "users" or other terms are a stand-in for a broader group of individuals, such as those suffering from depression. When algorithms detect depression in people who use Twitter (hence "user"), the approach may only be applicable to this group, not to non-users of Twitter [15]. Imprecise language describing users to be *all* individuals with depression can harm reproducibility, especially when these approaches imply transferability across disciplines and into practice (*e.g.*, algorithms built on Twitter users deployed among clinical patients at large). We envision this issue emerging in other areas of HCML, where stakeholders may anticipate generalizability of HCML "solutions" to practice that are not correctly described in the papers.

6.2.2 Consequences to Health and Communities. The representations we identify in our Findings risk diminishing the importance of the context of health and related communities. Context is a key attribute within systems and interaction design, described by Dourish as emergent, everyday, and essential [63]. By translating away from the complexities of experiences in this dataset for a more compact or compressed mathematical representation, HCML will lose fidelity in the experiences of those struggling with mental health. What could be the consequences of this loss of context?

Consequences to Mental Health: Algorithmic representations and abstractions necessarily compress the complexity of mental illness and well-being; although necessary for generalizing, this can cause downstream impacts on understanding the unique experiences and symptomatology of mental illness. When researchers use binary classification, this reduces mental health status to two corresponding machine learning classes "positive"/"negative" and the corresponding Disorder-focused "suffering"/"not suffering" can erase subtleties around the spectrum of mental illness

severity and co-morbid diagnoses [36]. We find additional evidence of this in our findings through the focus on "posts" as singular moments of distress, where it is not clear that a post is enough context to evaluate and determine risk.

By abstracting away from complexities around diagnosis, we may diminish the predictive power and application of these technologies to clinical scenarios. Patient-centered turns in medicine have sought to bring more experiential, contextual, and personal information to the process of providing medical care [18, 181]. Feuston and Piper highlighted this recently, focusing on how "small stories" of mental health expressed through social media contextualized individuals' experiences [69]. Through the necessity of creating blunt abstractions, research may oversimplify the experiences of mental illness as more than just posts or diagnostic status.

Consequences to Online Health Communities: These translations complicate the role of online communities in driving conversations on mental health. The benefits of online community participation for mental health are well-documented — they provide support and belonging, empowerment in light of stigma, and gateways to knowledge and better health [22, 96, 130].

The goals and practices of the community may not necessarily align with the goals of the HCML researcher, as data is translated out of the context it may not have been envisioned or desired to be used for [95, 124]. This relates to Nissenbaum's notion of contextual integrity [133], where expectations of privacy are necessarily situated in the context which information is shared. In taking this data for alternative purposes, research may distort reasonable expectations of privacy and acceptable behavior in these communities, overstepping moral boundaries with these communities [36]. On the other hand, overindexing to a focus on the community by researchers can be burdensome for the community of interest, and may also cause harm in publication by deanonymizing or identifying individuals and communities, making them a potential target. Without engaging the context of each community, it is difficult to know how to frame the role of the community in understanding mental health discussions.

Overall, we worry this work could be disempowering for the individuals and communities whose data is used for this research. This is in part because there is little relationship between them and HCML researchers and little reflection on how HCML work may legitimately lead to empowerment [155]. A human-centered paradigm should give agency to individuals and communities who technology is designed for by involving their perspective [105] – but in what ways should this occur in HCML? What are best practices for promoting empowerment for communities [155]? It is an open question how this should be conducted given the risks and balances above, especially as the data from these communities is transformed for secondary and often unexpected purposes.

6.2.3 Consequences to Individuals. Discourse can diminish as well as promote the validity of identities of vulnerable populations [33, 64, 129], and can even reproduce sexist gender roles [120]. Drawing on these insights, we highlight some potential consequences directed at individuals as a result of the representations of the human.

Increasing Stigma: Discourse can contribute to *stigma*, an attribute that makes an individual undesirable, tainted, or socially unworthy [4, 78]. Stigma has very dangerous consequences for those who suffer from mental disorders, such as causing delays in, non-compliance with, or unwillingness to participate in treatment [49], diminished social support [146], and decreased self-esteem and self-efficacy [146]. For other stigmatized identities, physical/tangible harms of stigma exist [86, 117].

We noticed the use of highly stigmatizing language choices. Some discourse compared those suffering from mental disorders to "normal," "regular," or "neurotypical" people. We were also concerned by discourse around sufferers, as language like this erases a person's complex identity – a person who suffers from schizophrenia – and replaces it with their mental disorder as the operative portion of their being – a "schizophrenic." These examples run counter to research and

journalistic/reporting guidelines on how to discuss mental disorders without promoting stigma [79, 152]. In particular, mental health advocates have moved away from using the term "neurotypical" or "normal" in favor of acknowledging individuals where they are [42]. We worry that discourse likely stigmatizes these identities and risks harming individuals in datasets that we as researchers intend to help. Furthermore, when outside audiences engage with this work, such as non-computer scientists, lay people, and journalists, these confusing representations can propagate outside of scholarly engagement and perpetuate standards within larger, publicly held conceptualizations [120].

Risking Dehumanization: The implications of research are decidedly human-centered – many documents celebrate impacts for monitoring, intervention efforts, and fundamental shifts in how we diagnose and treat mental disorders. Yet, our discourse analysis points to other forms of engagement with people as discursive objects. At the extreme, humans become the literal objects in social media: "accounts" or "blogs," and the data objects themselves, "positive/negative" and "samples," "extracted" or "harvested" for value.

What are the risks of depersonalization and dehumanization? As D'Ignazio and Klein contend, "Without the ability of individuals and communities to shape the terms of their own data collection, their bodies can be mined and their data can be extracted far too easily – and done so by powerful institutions who rarely have their best interests at heart." [59] Dehumanizing data can lead to researcher negligence, ignoring risks in algorithmic design and practice because the humans have been, in Haslam's terms, made "psychologically distant" [85]. These algorithmic harms have been recognized in other areas [103], and similar harms may occur in part due to this research.

Dehumanization clouds the responsibilities and ethical priorities of researchers. When the human is translated into abstracted representations, it may be easier to justify certain ethical or methods decisions. These risks could include revealing personal or private data, failing to deanonymize quotes [6], releasing datasets that were unethically gathered [142], and conducting experiments on individuals without their consent [107]. It is not only in this case study that such concerns have been articulated—we see such tensions emerge in critical data studies, which has challenged conceptualizing who the human research subject is in social media scenarios [123, 187] and how researchers interpret their own responsibilities [174].

Outside of these impacts, powerful actors with conflicting interests could cause harm to individuals. We worry that algorithms will reproduce discriminatory outcomes that perpetuate societal injustice towards those least able to counter the harmful effects of algorithmic inference [9, 134]. For mental health, this may mean discriminatory outcomes for those already struggling with a stigmatized illness. By depersonalizing and dehumanizing the individuals in HCML, CS researchers may develop systems that do not meet people's needs and desires, and may reproduce socially unjust and undesirable outcomes.

6.3 Implications and Guidelines for HCML

Our work provides an opportunity to reflect on practices in a specific case of HCML early in its history. Explicit focus on the practices within fields has been valuable to identify methodological problems and trends [67, 112], and redirecting towards more productive practices [17, 56].

In light of our findings, one may assume that there is a "correct" discourse to discuss these representations, and that by simply solving language, we will therefore avoid these consequences. This proposition is alluring, since it provides a simple and elegant solution to avoid the risks we identify above. However, directing deliberate language use is reductionist and dangerous for advocating meaningful engagement with these issues. It reduces a complex discursive representation to a checklist, an approach eschewed by ethicists who encourage researchers to adopt "ethics as a value" and process [35]. Rigid rules can segment and "bureaucratize" knowledge to only certain

stakeholders, in this case, to CSCW and HCI [111], running counter to a human-centered agenda. It may make adoption more difficult in other disciplines, as such a set of rules may feel as if other disciplines are left out [27].

We also do not want to suggest that ML and related computational approaches should be avoided and abandoned entirely. This is also reductionist, as abstraction through representation in mathematics is an important tool within scientific research. Our goals are to envision ways that HCML could work to recenter the human through its practices and make its commitment to humans clear [8]. We are cautiously optimistic that HCML can be more correctly oriented to humanistic representations within its research and align towards a stronger human-centered agenda. Our optimism is inspired by recent self-criticisms from practitioners within ML around similarly challenging problems [73, 92, 118, 158].

In this spirit, rather than advocate for rigid rules prescribed to others or an outright dismissal of the field, we offer a beginning set of guidelines informed by our findings for HCML. These guidelines are informed in part by our case study here and recent discussion around these issues [1, 145]. We intend that these guidelines start a conversation around these representations within HCML and the values of the community, rather than a prescribed "one-size-fits-all" solution to solve these challenging issues of representation.

Committing to Reproducibility in Descriptions. Given the challenges of interdisciplinarity and the risks to scientific practice we identify above, researchers and practitioners must be vigilant in their reporting practices within HCML.

This involves making it abundantly clear what proxy language is chosen and how it is operationalized, crucial to communicating the work to HCML's wide interdisciplinary audiences. One strong example of proxy language came from our corpus, where the authors identify how they refer to individuals: "The website collected the responses to a questionnaire to evaluate the degree of depression of the *Twitter users* who participated (hereinafter, *the participants*) and to collect the histories of participants activities on Twitter" [168][p. 3189]. Explicit proxy language may also make explicit how the transformations of data produced by individuals is being incorporated into the analysis. In another strong example of clear language, proxy language explains the machine learning transformations: "each *post* thus gave rise to a vector consisting of 93 input attributes and 1 label, or output attribute. The collection of all of these 459 vectors makes up the training data for our machine learning approach". [164][p. 509]. These examples demonstrate one approach to explicating the transformations of individuals into data to diminish objectification.

Another practice for reproducibility is that transitions between sections that necessitate proxy language need to be handled with care. This prevents necessary language for ML and data precision from propagating to the rest of the paper and potentially dehumanizing individuals. Good proxy representations make meaning explicit for researchers inside and outside the domain of publication, as well as to non-academic professionals interested in applying these findings. These practices and standards for reproducibility are not just the responsibility of the authors – reviewers and disciplinary communities reviewing HCML research can advocate for higher reproducibility standards and methods reporting within the community.

Collaborate with Interdisciplinary Experts to Build a "Shared Vocabulary." Bracken and Oughten argue that "interdisciplinary projects must allocate time to the development of shared vocabularies and understandings"[27][p. 371]. To resolve this, we encourage researchers to partner with domain experts through project collaborations. For our case study, key domain experts may be in-practice clinicians, medical doctors and researchers, or social workers with expertise in mental health. These partnerships are essential in bringing deep knowledge from other fields to

computational research, as well as experiences that will shape a more focused and human-centered agenda. We imagine partnerships like these as a crucial component of HCML research more broadly.

We also advocate for interdisciplinary workshops and shared spaces to develop community-wide shared understandings for this field. These interdisciplinary workshops are beginning to appear, and we are hopeful that future work between researchers from ML, HCI, NLP and other areas can come together with domain experts outside of CS to work on these problems. We hope that shared workshops and collaborative opportunities may additionally iterate on our guidelines here and refine an agenda for HCML situated between disciplines larger than just CS [1].

Be Mindful of Risky Practices Within Research. Language provides insights into the practices and relationships within society [74, 75], and we use this opportunity to reflect on what the discourse in these papers may indicate about practices behind HCML.

We encourage researchers to be mindful of practices within their work that may be harmful to individuals. Related directly to our findings, stigmatizing language harms people and communities [49, 117], and we generally advocate against its use. Journalistic outlets and non-profit organizations have created detailed standards on language that avoids reproducing or increasing stigma [79, 152]. These best practices in for language in mental health complement other topical examples of vulnerable or historically marginalized population that are of interest to HCML.

We also encourage practitioners to examine the entire research process to identify opportunities for engagement with representations of the human in practice. Our analysis shows that these risks could often happen through the abstraction steps necessary for data analysis. In addition to language, these harms may occur through other methods decisions, such as automatic gender recognition technologies that erase non-binary identities [103] or poor performance of language analysis on women and minorities [24]. We worry about data collection and reporting practices in papers that may expose information about people around sensitive and stigmatizing life experiences [6]. These questions of harm to the individuals should be adopted and navigated throughout the whole research process. Although some of these commitments exist out of examining language and discourse directly, we argue that better practices throughout the HCML research process will encourage better representations.

Put the Human Back into HCML. Human-centered methods will direct better engagement with these communities, and determine the appropriate modes of representation. These include participatory approaches [14], interviews and field studies, and collaborations with those who best know the topic domain. We envision collaborative opportunities for HCML to work alongside these participatory approaches to understand individuals' behaviors, aspirations, and goals – providing insight into the context lost through mathematical translations. These partnerships can also be key to understanding how HCML can be used to provide agency, power, and empowerment back to the communities in scientifically responsible ways.

It is important to recognize that engagement with these guidelines is only one portion of a larger paradigm of human-centered research. Following these steps does not guarantee that a researcher or a project is "human-centered" and therefore safe to conduct or publish, nor do they absolve the researcher of representational issues or harms caused by algorithms. Further, our guidelines are inspired by and contribute to a holistic argument for human-centeredness within this research. Human-centered machine learning is more than just an approach to AI, ML, or other disciplines that brings "humans in the loop" or solely offer methods innovations. Human-centeredness is a deliberate refocus on the needs of humans, communities, and society and identifying the appropriate tools to solve problems [8, 105]. Commitment to this process is difficult – it involves collaborations with stakeholders and domain experts, an investment in the needs and goals of a particular problem domain, and then (potentially) engineering a solution that brings these people along as equitable

partners. In fact, by following these approaches, these processes may reveal that an HCML solution is not appropriate or desirable for a problem domain [16, 158].

In sum, we strongly advocate for HCML researchers to consider these approaches, in part to avoid the risks of "big data hubris" [112], as well as to do right by the communities and individuals our work claims to help.

7 CONCLUSION

In this paper, we conduct a discourse analysis to understand the practices of representation within human-centered machine learning (HCML). We identify a dataset of 55 interdisciplinary papers from the case of predicting mental health status on social media data. Our results suggest that competing discourses interact throughout to conceptualize and give agency to the humans within this dataset. Our findings show how these five discourses create paradoxical subject and object representations of the human, which may inadvertently risk dehumanization. We have demonstrated how these competing discourses cause harms to scientific rigor and reproducibility, to understanding context in mental health and online communities, and to the individuals who are the beneficiaries of these analyses. We look forward to similar analyses and reflections on other cases within HCML, and we are excited for future work from these perspectives as well as interdisciplinary critique.

ACKNOWLEDGMENTS

The authors would like to thank Brianna Dym, Michaelanne Dye, Os Keyes, Karen Boyd, and the anonymous reviewers for their feedback. This project was funded in part by NIH Award #R01MH117172 and NSF Award #1816403.

REFERENCES

[1] Ali Alkhatib. 2019. Anthropological/Artificial Intelligence & the HAI. https://ali-alkhatib.com/blog/anthropological-intelligence

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, Eric Horvitz Microsoft, Paul G Allen, and Adam Four-ney. 2019. Guidelines for Human-AI Interaction. *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019)* (2019), 13. https://doi.org/10.1145/3290605.3300233

[3] Cecilia Aragon, Clayton Hutto, Andy Echenique, Brittany Fiore-Gartland, Yun Huang, Jinyoung Kim, Gina Neff, Wanli Xing, and Joseph Bayer. 2016. Developing a research agenda for human-centered data science. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*. ACM, 529–535.

[4] Julio Arboleda-Flórez and Heather Stuart. 2012. From sin to science: Fighting the stigmatization of mental illnesses. *Canadian Journal of Psychiatry* 57, 8 (2012), 457–463. https://doi.org/10.1177/070674371205700803

[5] American Psychiatric Association et al. 2013. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

[6] John W. Ayers, Theodore L. Caputi, Camille Nebeker, and Mark Dredze. 2018. Don't quote me: reverse identification of research participants in social media studies. *npj Digital Medicine* 1, 1 (2018), 30. https://doi.org/10.1038/s41746-018-0036-2

[7] L. Bannon. 1991. From Human Factors to Human Actors: The Role of Psychology and Human-Computer Interaction Studies in System Design. In *Design at Work: Cooperative Design of Computer Systems*. 25–44.

[8] Liam Bannon. 2011. Reimagining HCI: toward a more human-centered perspective. *interactions* 18, 4 (2011), 50–57.

[9] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.

[10] Andrew Barry, Georgina Born, and Gisa Weszkalnys. 2008. Logics of interdisciplinarity. *Economy and Society* 37, 1 (2008), 20–49.

[11] Henry H Bauer. 1990. Barriers against interdisciplinarity: implications for studies of science, technology, and society (STS. *Science, Technology, & Human Values* 15, 1 (1990), 105–119.

[12] Eric P. S. Baumer. 2015. Reflective Informatics: Conceptual Dimensions for Designing Technologies of Reflection. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. ACM, Seoul, 585–594. https://doi.org/10.1145/2702123.2702234

[13] Eric P. S. Baumer. 2015. Usees. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. ACM Press, Seoul, 3295–3298. https://doi.org/10.1145/2702123.2702147
[14] Eric P. S. Baumer. 2017. Toward Human-Centered Algorithm Design. *Big Data & Society* 4, 2 (2017). https://doi.org/10.1177/2053951717718854
[15] Eric P. S. Baumer and Jed R. Brubaker. 2017. Post-Userism. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. ACM, Denver, CO, 6291–6303. https://doi.org/10.1145/3025453.3025740
[16] Eric P. S. Baumer and M. Six Silberman. 2011. When the Implication Is Not to Design (Technology). In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. Vancouver, BC, Canada, 2271–2274. https://doi.org/10.1145/1978942.1979275
[17] Genevieve Bell and Paul Dourish. 2007. Yesterday's Tomorrows: Notes on Ubiquitous Computing's Dominant Vision. *Personal and Ubiquitous Computing* 11, 2 (Jan. 2007), 133–143. https://doi.org/10.1007/s00779-006-0071-x
[18] Jozien Bensing. 2000. Bridging the gap.: The separate worlds of evidence-based medicine and patient-centered medicine. *Patient education and counseling* 39, 1 (2000), 17–25.
[19] Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. 94–102.
[20] Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multitask learning for mental health conditions with limited social media data. *EACL* (2017). http://www.aclweb.org/anthology/E17-1015
[21] Roni Berger. 2015. Now I see it, now I don't: Researcher's position and reflexivity in qualitative research. *Qualitative research* 15, 2 (2015), 219–234.
[22] Natalie Berry, Fiona Lobban, Maksim Belousov, Richard Emsley, Goran Nenadic, and Sandra Bucci. 2017. #WhyWeTweetMH: understanding why people use Twitter to discuss mental health problems. *Journal of medical Internet research* 19, 4 (2017), e107.
[23] Michael L Birnbaum, Sindhu Kiranmai Ernala, Asra F Rizvi, Munmun De Choudhury, and John M Kane. 2017. A Collaborative Approach to Identifying Social Media Markers of Schizophrenia by Employing Machine Learning and Clinical Appraisals. *JMIR* 19, 8 (aug 2017).
[24] Su Lin Blodgett and Brendan O'Connor. 2017. Racial disparity in natural language processing: A case study of social media african-american english. *arXiv preprint arXiv:1707.00061* (2017).
[25] danah boyd and Kate Crawford. 2012. Critical Questions for Big Data. *Information, Communication & Society* 15, 5 (2012), 662–679. https://doi.org/10.1080/1369118X.2012.678878
[26] danah boyd and Nicole B Ellison. 2007. Social network sites: Definition, history, and scholarship. *Journal of computer-mediated Communication* 13, 1 (2007), 210–230.
[27] Louise J Bracken and Elizabeth A Oughton. 2006. "What do you mean?" The importance of language in developing interdisciplinary research. *Transactions of the Institute of British Geographers* 31, 3 (2006), 371–382.
[28] Scott R Braithwaite, Christophe Giraud-Carrier, Josh West, Michael D Barnes, and Carl Lee Hanson. 2016. Validating Machine Learning Algorithms for Twitter Data Against Established Measures of Suicidality. *JMIR Mental Health* 3, 2 (2016), e21.
[29] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.
[30] Jed R. Brubaker and Gillian R. Hayes. 2011. SELECT * FROM USER: Infrastructure and Socio-Technical Representation. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW)*. ACM, Hangzhou, China, 369–378. https://doi.org/10.1145/1958824.1958881
[31] Pete Burnap, Walter Colombo, and Jonathan Scourfield. 2015. Machine Classification and Analysis of Suicide-Related Communication on Twitter. In *HT (HT '15)*. ACM, 75–84.
[32] Jenna Burrell. 2016. How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms. *Big Data & Society* 3, 1 (Jan. 2016), 2053951715622512. https://doi.org/10.1177/2053951715622512
[33] Judith Butler. 2001. Doing Justice to Someone: Sex Reassignment and Allegories of Transsexuality. *GLQ: A Journal of Lesbian and Gay Studies* 7, 4 (Sept. 2001), 621–636.
[34] Michel Callon. 1986. Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of Saint Brieuc Bay. In *Power, Action and Belief: A New Sociology of Knowledge?*, John Law (Ed.). Number 32 in Sociological Review Monograph. Routledge, London, 196–223.
[35] Katherine Carpenter and David Dittrich. 2011. Bridging the distance: removing the technology buffer and seeking consistent ethical analysis in computer security research. In *1st International Digital Ethics Symposium. Loyola University Chicago Center for Digital Ethics and Policy.*
[36] Stevie Chancellor, Michael Birnbaum, Eric Caine, Vincent Silenzio, and Munmun De Choudhury. 2019. A Taxonomy of Ethical Tensions in Inferring Mental Health States from Social Media. In *Proceedings of the 2019 FAT* Conference - Fairness, Accountability and Transparency*. ACM.

[37] Stevie Chancellor, Zhiyuan (Jerry) Jerry Lin, Erica L. Goodman, Stephanie Zerwas, and Munmun De Choudhury. 2016. Quantifying and Predicting Mental Illness Severity in Online Pro-Eating Disorder Communities. *CSCW*, 1169–1182. http://dl.acm.org/citation.cfm?doid=2818048.2819973

[38] Stevie Chancellor, Tanushree Mitra, and Munmun De Choudhury. 2016. Recovery Amid Pro-Anorexia: Analysis of Recovery in Social Media. *CHI*, 2111–2123.

[39] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 32.

[40] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. The bag of communities: Identifying abusive behavior online with preexisting Internet data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 3175–3187.

[41] Qijin Cheng, Tim Mh H Li, Chi-Leung Leung Kwok, Tingshao Zhu, and Paul Sf F Yip. 2017. Assessing suicide risk and emotional distress in Chinese social media: A text mining and machine learning study. *Journal of Medical Internet Research* 19, 7 (jul 2017), 1–10.

[42] John A Clausen. 1981. Stigma and mental disorder: Phenomena and terminology. *Psychiatry* 44, 4 (1981), 287–296.

[43] Mike Conway. 2014. Ethical issues in using twitter for public health surveillance and research: Developing a taxonomy of ethical concepts from the research literature. *Journal of Medical Internet Research* 16, 12 (2014).

[44] Geoff Cooper and John Bowers. 1995. Representing the User: Notes on the Disciplinary Rhetoric of HCI. In *The Social and Interactional Dimensions of Human-Computer Interfaces*, Peter J. Thomas (Ed.). Cambridge University Press, Cambridge, 48–66.

[45] Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying Mental Health Signals in Twitter. *CLPsych* 2014, 51–60.

[46] Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *CLPsych*. 1–10.

[47] Glen Coppersmith, Craig Harman, and Mark H Dredze. 2014. Measuring post traumatic stress disorder in Twitter. In *ICWSM*, Vol. 2. 579–582.

[48] Glen Coppersmith, Kim Ngo, Ryan Leary, and Anthony Wood. 2016. Exploratory analysis of social media prior to a suicide attempt. In *CLPsych*. 106–117.

[49] Patrick Corrigan. 2004. How stigma interferes with mental health care. *American psychologist* 59, 7 (2004), 614.

[50] Munmun De Choudhury. 2015. Anorexia on Tumblr : A Characterization Study on Anorexia. In *Proceedings of DH'15: 5th ACM Digital Health Conference. DH'15. (DH '15)*. ACM, 43–50.

[51] Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Predicting postpartum changes in emotion and behavior via social media. *CHI*, 3267–3276.

[52] Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Social Media As a Measurement Tool of Depression in Populations. In *WebSci (WebSci '13)*. ACM, 47–56.

[53] Munmun De Choudhury, Scott Counts, Eric J. Horvitz, and Aaron Hoff. 2014. Characterizing and Predicting Postpartum Depression from Shared Facebook Data. In *CSCW (CSCW '14)*. ACM, 626–638.

[54] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. *ICWSM* 2, 128–137.

[55] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *CHI*. ACM, 2098–2110. https://doi.org/10.1145/2858036.2858207

[56] Nicola Dell and Neha Kumar. 2016. The ins and outs of HCI for development. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2220–2232.

[57] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[58] Michael A. DeVito, Ashley Marie Walker, and Jeremy Birnholtz. 2018. 'Too Gay for Facebook': Presenting LGBTQ+ Identity Throughout the Personal Social Media Ecosystem. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (Nov. 2018), 44:1–44:23. https://doi.org/10.1145/3274313

[59] Catherine D'Ignazio and Lauren Klein. 2019. *Chapter One: Bring Back the Bodies*. MIT Open Press. https://bookbook.pubpub.org/pub/zrlj0jqb

[60] Tawanna R Dillahunt, Xinyi Wang, Earnest Wheeler, Hao Fei Cheng, Brent J Hecht, and Haiyi Zhu. 2017. The Sharing Economy in Computing: A Systematic Literature Review. *PACMHCI* 1, CSCW (2017), 38–1.

[61] Carl DiSalvo, Phoebe Sengers, and Hrönn Brynjarsdóttir. 2010. Mapping the Landscape of Sustainable HCI. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. ACM, Atlanta, GA, 1975–1984.

[62] Derek Doran, Sarah Schulz, and Tarek R Besold. 2017. What does explainable AI really mean? A new conceptualization of perspectives. *arXiv preprint arXiv:1710.00794* (2017).

[63] Paul Dourish. 2004. What we talk about when we talk about context. *Personal and ubiquitous computing* 8, 1 (2004), 19–30.

[64] Jack Drescher, Peggy Cohen-Kettenis, and Sam Winter. 2012. Minding the body: Situating gender identity diagnoses in the ICD-11. *International Review of Psychiatry* 24, 6 (2012), 568–577.

[65] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. ACM, 214–226.

[66] Ronald M Epstein and Richard L Street. 2011. The values and value of patient-centered care.

[67] Sindhu Kiranmai Ernala, Michael L Birnbaum, Kristin A Candan, Asra F. Rizvi, William A Sterling, John M Kane, and Munmun De Choudhury. 2019. Methodological Gaps in Predicting Mental Health States from Social Media : Triangulating Diagnostic Signals. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*.

[68] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 259–268.

[69] Jessica L Feuston and Anne Marie Piper. 2019. Everyday Experiences Small Stories and Mental Illness on Instagram. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 265.

[70] Casey Fiesler and Blake Hallinan. 2018. "We Are the Product": Public Reactions to Online Data Sharing and Privacy Controversies in the Media. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. ACM, Montréal, QC, 53:1–53:13. https://doi.org/10.1145/3173574.3173627

[71] Casey Fiesler and Nicholas Proferes. 2018. "Participant" Perceptions of Twitter Research Ethics. *Social Media+ Society* 4, 1 (2018).

[72] National Commission for the Protection of Human Subjects of Biomedicaland Behavioral Research. 1978. *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research*. Superintendent of Documents.

[73] Jessica Zosa Forde and Michela Paganini. 2019. The Scientific Method in the Science of Machine Learning. *arXiv preprint arXiv:1904.10922* (2019).

[74] Michel Foucault. 1972. *The Archaeology of Knowledge*. Pantheon Books, New York.

[75] James Paul Gee. 2011. *An introduction to discourse analysis: Theory and method*. Routledge.

[76] Marco Gillies, Rebecca Fiebrink, Atau Tanaka, Jérémie Garcia, Frédéric Bevilacqua, Alexis Heloir, Fabrizio Nunnari, Wendy Mackay, Saleema Amershi, Bongshin Lee, et al. 2016. Human-centred machine learning. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 3558–3565.

[77] George Gkotsis, Anika Oellrich, Sumithra Velupillai, Maria Liakata, Tim J P Hubbard, Richard J B Dobson, and Rina Dutta. 2017. Characterisation of mental health conditions in social media using Informed Deep Learning. *SCIENTIFIC REPORTS* 7 (mar 2017). https://doi.org/10.1038/srep45141

[78] Erving Goffman. 2009. *Stigma: Notes on the management of spoiled identity*. Simon and Schuster.

[79] Laura Greenstein. [n. d.]. Why Suicide Reporting Guidelines Matter. https://www.nami.org/Blogs/NAMI-Blog/June-2018/Why-Suicide-Reporting-Guidelines-Matter

[80] Jonathan Grudin. 2009. AI and HCI: Two Fields Divided by a Common Focus. *AI Magazine* 30, 4 (Sept. 2009), 48–48. https://doi.org/10.1609/aimag.v30i4.2271

[81] Li Guan, Bibo Hao, Qijin Cheng, Paul SF F Yip, and Tingshao Zhu. 2015. Identifying Chinese Microblog Users With High Suicide Probability Using Internet-Based Profile and Linguistic Features: Classification Model. *JMIR Mental Health* 2, 2 (2015), e17. http://mental.jmir.org/2015/2/e17/

[82] Björn Hammarfelt, Fredrik Åström, and Joacim Hansson. 2017. Scientific publications as boundary objects: theorising the intersection of classification and research evaluation. In *Information research*, Vol. 22.

[83] Ellie Harmon and Melissa Mazmanian. 2013. Stories of the Smartphone in everyday discourse: conflict, tension & instability. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 1051–1060.

[84] Anne-Wil Harzing et al. 2007. Publish or perish. (2007).

[85] Nick Haslam. 2006. Dehumanization: An integrative review. *Personality and social psychology review* 10, 3 (2006), 252–264.

[86] Mark L Hatzenbuehler, Anna Bellatorre, Yeonjin Lee, Brian K Finch, Peter Muennig, and Kevin Fiscella. 2014. Structural stigma and all-cause mortality in sexual minority populations. *Social Science & Medicine* 103 (2014), 33–41.

[87] Anna Lauren Hoffmann. 2019. Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication, and Society* (2019).

[88] Anna Lauren Hoffmann, Nicholas Proferes, and Michael Zimmer. 2018. "Making the world more open and connected": Mark Zuckerberg and the discursive construction of Facebook and its users. *New Media and Society* 20, 1 (2018), 199–218. https://doi.org/10.1177/1461444816660784

[89] Jake M Hofman, Amit Sharma, and Duncan J Watts. 2017. Prediction and explanation in social systems. *Science* 355, 6324 (2017), 486–488.

[90] Christopher M Homan, Ravdeep Johar, Tong Liu, Megan Lytle, Vincent Silenzio, and Cecilia Ovesdotter Alm. 2014. Toward Macro-Insights for Suicide Prevention: Analyzing Fine-Grained Distress at Scale. In *CLPsych*. 107.

[91] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 159–166.

[92] Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2. 591–598.

[93] Xiaolei Huang, Xin Li, Lei Zhang, Tianli Liu, David Chiu, and Tingshao Zhu. 2015. Topic Model for Identifying Suicidal Ideation in Chinese Microblog. In *29th Pacific Asia Conference on Language, Information and Computation*. Proceeedings of the 29th Pacific Asia Conference on Language, 553–562. http://www.aclweb.org/anthology/Y15-1064

[94] Xiaolei Huang, Lei Zhang, David Chiu, Tianli Liu, Xin Li, and Tingshao Zhu. 2014. Detecting Suicidal Ideation in Chinese Microblogs with Psychological Lexicons. *2014 IEEE International Conference on Autonomic and Trusted Computing, 2014 IEEE International Conference on Scalable Computing and Communications and Associated Symposia/Workshops, UIC-ATC-ScalCom 2014* 2014 (2014), 844–849.

[95] James M Hudson and Amy Bruckman. 2004. "Go away": participant objections to being studied and the ethics of chatroom research. *The Information Society* 20, 2 (2004), 127–139.

[96] Jina Huh and Mark S Ackerman. 2012. Collaborative help in chronic disease management: supporting individualized problems. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 853–862.

[97] Luke Hutton and Tristan Henderson. 2015. " I didn't sign up for this!": Informed consent in social network research. In *Ninth International AAAI Conference on Web and Social Media*.

[98] Kori Inkpen, Munmun De Choudhury, Stevie Chancellor, Michael Veale, and Eric P.S. Baumer. 2019. Where is the Human? Bridging the Gap Between AI and HCI. In *2019 CHI Extended Abstracts*. ACM.

[99] Lilly C. Irani and M. Six Silberman. 2016. Stories We Tell About Labor: Turkopticon and the Trouble with "Design". In *Proceedings of CHI 2016*. 4573–4586.

[100] Alejandro Jaimes, Daniel Gatica-Perez, Nicu Sebe, and Thomas S Huang. 2007. Guest Editors' Introduction: Human-Centered Computing–Toward a Human Revolution. *Computer* 40, 5 (2007), 30–34.

[101] Zunaira Jamil, Diana Inkpen, Prasadith Buddhitha, and Kenton White. 2017. Monitoring Tweets for Depression to Detect At-risk Users. In *CLPsych*. 32–40.

[102] Gopinaath Kannabiran, Jeffrey Bardzell, and Shaowen Bardzell. 2011. How HCI talks about sexuality: discursive strategies, blind spots, and opportunities for future research. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 695–704.

[103] Os Keyes. 2018. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 88.

[104] Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 1301–1318.

[105] Rob Kling and Susan Leigh Star. 1998. Human centered systems in the perspective of organizational and social informatics. *ACM SIGCAS Computers and Society* 28, 1 (1998), 22–29.

[106] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 1885–1894.

[107] Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. 2014. Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks. *Proceedings of the National Academy of Sciences (PNAS)* 111, 24 (2014), 8788–8790.

[108] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 1675–1684.

[109] Roberta Lamb and Rob Kling. 2003. Reconceptualizing users as social actors in information systems research. *MIS quarterly* (2003), 197–236.

[110] Bruno Latour. 1993. Ethnography of a "High-Tech" Case: About Aramis. In *Technological Choices: Transformations in Material Culture since the Neolithic*, Pierre Lemonnier (Ed.). Routledge, London, 372–398.

[111] Roderick J Lawrence and Carole Després. 2004. Futures of transdisciplinarity. *Futures* 4, 36 (2004), 397–405.

[112] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. The Parable of Google Flu: Traps in Big Data Analysis. *Science* 343, 6167 (2014), 1203–1205.

[113] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. 2009. Computational Social Science. *Science* 323, 5915 (Feb. 2009), 721–723.

[114] Alessandro Liberati, Douglas G. Altman, Jennifer Tetzlaff, Cynthia Mulrow, Peter C. Gøtzsche, John P.A. Ioannidis, Mike Clarke, P. J. Devereaux, Jos Kleijnen, and David Moher. 2009. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *PLoS*

Medicine 6, 7 (2009). https://doi.org/10.1371/journal.pmed.1000100

[115] Huijie Lin, Jia Jia, Quan Guo, Yuanyuan Xue, Qi Li, Jie Huang, Lianhong Cai, and Ling Feng. 2014. User-level psychological stress detection from social media using deep neural network. In *MM*. IEEE, New York, NY, USA, 507–516.

[116] H Lin, J Jia, L Nie, G Shen, and T S Chua. 2016. What Does Social Media Say about Your Stress?. *IJCAI* (2016). http://www.ijcai.org/Proceedings/16/Papers/531.pdf

[117] Bruce G Link and Jo C Phelan. 2006. Stigma and its public health implications. *The Lancet* 367, 9509 (2006), 528–529.

[118] Zachary C Lipton and Jacob Steinhardt. 2018. Troubling trends in machine learning scholarship. *arXiv preprint arXiv:1807.03341* (2018).

[119] Kate Loveys, Patrick Crutchley, Emily Wyatt, and Glen Coppersmith. 2017. Small but Mighty: Affective Micropatterns for Quantifying Mental Health from Social Media Language. In *CLPsych*. 85–95.

[120] Emily Martin. 1991. The egg and the sperm: How science has constructed a romance based on stereotypical male-female roles. *Signs: Journal of Women in Culture and Society* 16, 3 (1991), 485–501.

[121] Naoki Masuda, Issei Kurahashi, and Hiroko Onari. 2013. Suicide Ideation of Individuals in Online Social Networks. *PLOS ONE* 8, 4 (apr 2013).

[122] K McManus, E K Mallory, R L Goldfeder, W A Haynes, and J D Tatum. 2015. Mining Twitter Data to Improve Detection of Schizophrenia. *AMIA* 2015 (2015), 122–126.

[123] Jacob Metcalf and Kate Crawford. 2016. Where are human subjects in Big Data research? The emerging ethics divide. *Big Data & Society* 3, 1 (2016), 205395171665021.

[124] Jude Mikal, Samantha Hurst, and Mike Conway. 2016. Ethical issues in using Twitter for population-level depression monitoring: a qualitative study. *BMC medical ethics* 17, 1 (2016), 22.

[125] Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 2015. Quantifying the language of schizophrenia in social media. In *CLPsych*. 11–20.

[126] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 220–229.

[127] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining explanations in AI. (2019), 279–288.

[128] T Nakamura, K Kubo, Y Usuda, and E Aramaki. 2014. Defining patients with depressive disorder by using textual information. *AAAI* (2014).

[129] Viviane Namaste. 2000. *Invisible lives: The erasure of transsexual and transgendered people*. University of Chicago Press.

[130] JA Naslund, KA Aschbrenner, LA Marsch, and SJ Bartels. 2016. The future of mental health care: peer-to-peer support and social media. *Epidemiology and psychiatric sciences* 25, 2 (2016), 113–122.

[131] Thin Nguyen, Dinh Phung, Bo Dao, Svetha Venkatesh, and Michael Berk. 2014. Affective and content analysis of online depression communities. *Ieee Transactions on Affective Computing* 5, 3 (2014), 217–226.

[132] Matthew C Nisbet and Chris Mooney. 2007. Framing science. *Science* 316, 5821 (2007), 56–56.

[133] Helen Nissenbaum. 2011. A Contextual Approach to Privacy Online. *Daedalus* 140, 4 (Sept. 2011), 32–48.

[134] Safiya Umoja Noble. 2018. *Algorithms of oppression: How search engines reinforce racism*. NYU Press.

[135] Bridianne O'Dea, Stephen Wan, Philip J. Batterham, Alison L. Calear, Cecile Paris, and Helen Christensen. 2015. Detecting suicidality on twitter. *Internet Interventions* 2, 2 (2015), 183–188.

[136] Sungkyu Park, Sang Won Lee, Jinah Kwak, Meeyoung Cha, and Bumseok Jeong. 2013. Activities on Facebook reveal the depressive state of users. *Journal of Medical Internet Research* 15, 10 (2013), 1–15.

[137] Chanda Phelan, Cliff Lampe, and Paul Resnick. 2016. It's Creepy, But It Doesn't Bother Me. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. ACM, San Jose, CA, 5240–5251.

[138] Daniel Preotiuc-Pietro, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, H Andrew Schwartz, and Lyle Ungar. 2015. The Role of Personality, Age and Gender in Tweeting about Mental Illnesses. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* January (2015), 21–30.

[139] Victor M Prieto, Sergio Matos, Manuel Alvarez, Fidel Cacheda, and Jose Luis Oliveira. 2014. Twitter: A Good Place to Detect Health Conditions. *PLOS One* 9, 1 (jan 2014).

[140] Andrew G Reece and Christopher M Danforth. 2017. Instagram photos reveal predictive markers of depression. *EPJ DATA SCIENCE* 6 (aug 2017), 1–34.

[141] Andrew G. Reece, Andrew J. Reagan, Katharina L.M. M Lix, Peter Sheridan Dodds, Christopher M. Danforth, and Ellen J. Langer. 2017. Forecasting the onset and course of mental illness with Twitter data. *SCIENTIFIC REPORTS* 7, 1 (oct 2017).

[142] Brian Resnick. 2016. Researchers just released profile data on 70,000 OkCupid users without permission. *Vox. Retreived from: https://www. vox. com/2016/5/12/11666116/70000-okcupid-users-data-release* (2016).

[143] Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-an Nguyen, and Jordan Boyd-Graber. 2015. Beyond LDA: exploring supervised topic modeling for depression-related language in Twitter. In *CLPsych*, Vol. 1. 99–107.

[144] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 1135–1144.

[145] Mark O Riedl. 2019. Human-Centered Artificial Intelligence and Machine Learning. *arXiv preprint arXiv:1901.11184* (2019).

[146] Nicolas Rüsch, Matthias C Angermeyer, and Patrick W Corrigan. 2005. Mental illness stigma: concepts, consequences, and initiatives to reduce stigma. *European psychiatry* 20, 8 (2005), 529–539.

[147] Koustuv Saha, Larry Chan, Kaya De Barbaro, Gregory D Abowd, and Munmun De Choudhury. 2017. Inferring Mood Instability on Social Media by Leveraging Ecological Momentary Assessments. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3 (sep 2017), 95:1—-95:27.

[148] Koustuv Saha and Munmun De Choudhury. 2017. Modeling Stress with Social Media Around Incidents of Gun Violence on College Campuses. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW (dec 2017), 92:1–92:27.

[149] Pedro Sanches, Pavel Karpashevich, Gavin Doherty, Axel Janson, Charles Windlin, Corina Sas, Camille Nadal, and Kristina Höök. 2019. HCI and Affective Health. Taking stock of a decade of studies and charting future research directions. *Conference on Human Factors in Computing Systems -CHI'19* (2019), In press.

[150] Elvis Saravia, Chun Hao Chang, Renaud Jollet De Lorenzo, and Yi Shin Chen. 2016. MIDAS: Mental illness detection and analysis via social media. In *ASONAM (ASONAM '16)*. IEEE Press, 1418–1421.

[151] Christine Satchell and Paul Dourish. 2009. Beyond the User: Use and Non-Use in HCI. In *Proceedings of the Australasian Computer-Human Interaction Conference (OZCHI)*. ACM, Melbourne, Australia, 9–16.

[152] Katia Savchuk. [n. d.]. 5 Tips for Journalists Covering Mental and Behavioral Health. https://niemanstoryboard.org/stories/5-tips-for-journalists-covering-mental-and-behavioral-health

[153] Ari Schlesinger, W. Keith Edwards, and Rebecca E. Grinter. 2017. Intersectional HCI: Engaging Identity Through Gender, Race, and Class. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. ACM, Denver, CO, 5412–5427. https://doi.org/10.1145/3025453.3025766

[154] Ari Schlesinger, Kenton P O'Hara, and Alex S Taylor. 2018. Let's Talk About Race: Identity, Chatbots, and AI. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 315.

[155] Hanna Schneider, Malin Eiband, Daniel Ullrich, and Andreas Butz. 2018. Empowerment in HCI-A Survey and Framework. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 244.

[156] H Andrew Schwartz, Johannes Eichstaedt, Margaret L Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. Towards assessing changes in degree of depression through facebook. In *CLPsych*. 118–125.

[157] Elizabeth M Seabrook, Bpsych Hons, Margaret L Kern, and Nikki S Rickard. 2016. Social Networking Sites, Depression, and Anxiety: A Systematic Review. *JMIR Mental Health* 3, 4 (2016), e50. https://doi.org/10.2196/mental.5842

[158] Andrew D Selbst, Sorelle Friedler, Suresh Venkatasubramanian, Janet Vertesi, et al. 2019. Fairness and Abstraction in Sociotechnical Systems. In *ACM Conference on Fairness, Accountability, and Transparency (FAT*)*.

[159] Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, and Wenwu Zhu. 2017. Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution.. In *IJCAI*. 3838–3844.

[160] Judy Hanwen Shen and Frank Rudzicz. 2017. Detecting anxiety on Reddit. In *CLPsych*. 58–65.

[161] Yu-chun Shen, Tsung-ting Kuo, I-ning Yeh, Tzu-ting Chen, and Shou-de Lin. 2013. Exploiting Temporal Information in a Two-Stage Classification Framework for Content-Based Depression. *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (2013), 276–288.

[162] Ben Shneiderman and Pattie Maes. 1997. Direct manipulation vs. interface agents. *interactions* 4, 6 (1997), 42–61.

[163] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter* 19, 1 (2017), 22–36.

[164] T Simms, C Ramstedt, M Rich, M Richards, T Martinez, and Christophe Giraud-Carrier. 2017. Detecting Cognitive Distortions Through Machine Learning Text Analytics. *ICHI* (2017). http://ieeexplore.ieee.org/abstract/document/8031202/

[165] Susan Leigh Star and James R. Griesemer. 1989. Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science* 19, 3 (Aug. 1989), 387–420. https://doi.org/10.1177/030631289019003001

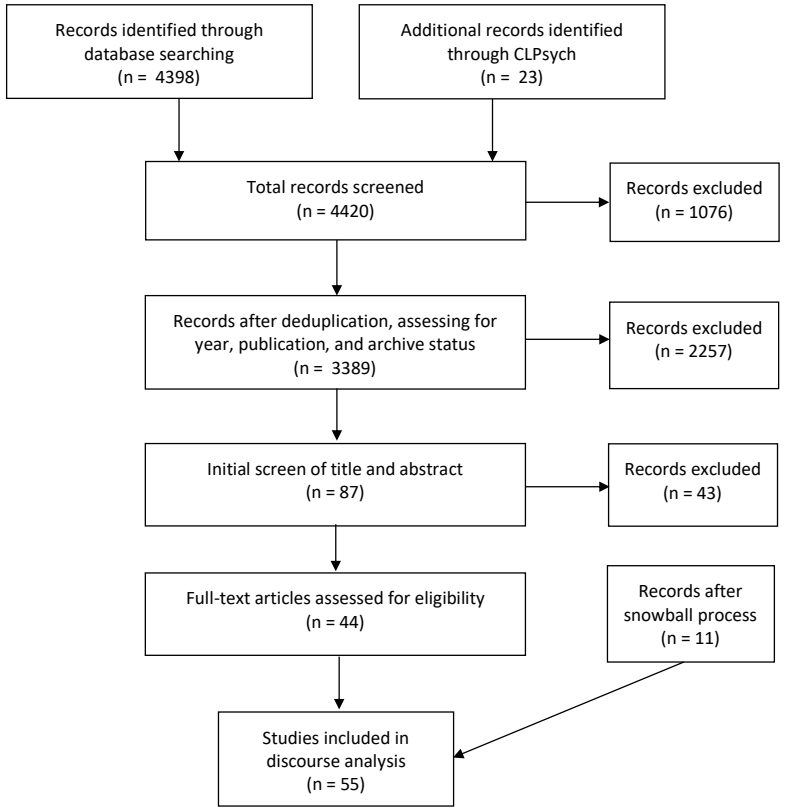
[166] Kate Starbird, Grace Muzny, and Leysia Palen. 2012. Learning from the crowd: collaborative filtering techniques for identifying on-the-ground Twitterers during mass disruptions. In *Proceedings of 9th International Conference on Information Systems for Crisis Response and Management, ISCRAM*. 1–10.

[167] Norman Makoto Su, Leslie S Liu, and Amanda Lazar. 2014. Mundanely miraculous: the robot in healthcare. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*. ACM, 391–400.
[168] Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. 2015. Recognizing Depression from Twitter Activity *(CHI '15)*. ACM, 3187–3196.
[169] Sho Tsugawa, Yukiko Mogi, Yusuke Kikuchi, Fumio Kishino, Kazuyuki Fujita, Yuichi Itoh, and Hiroyuki Ohsaki. 2013. On estimating depressive tendencies of twitter users utilizing their tweet data. In *2013 IEEE Virtual Reality (VR)*. IEEE, 1–4.
[170] Zeynep Tufekci. 2014. Engineering the Public: Big Data, Surveillance and Computational Politics. *First Monday* 19, 7 (2014).
[171] Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Fourth international AAAI conference on weblogs and social media*.
[172] Theo Van Leeuwen. 2005. Three models of interdisciplinarity. *A new agenda in (critical) discourse analysis: Theory, methodology and interdisciplinarity* (2005), 3–18.
[173] Nikhita Vedula and Srinivasan Parthasarathy. 2017. Emotional and Linguistic Cues of Depression from Social Media. In *DH*. ACM, 127–136. http://doi.acm.org/10.1145/3079452.3079465
[174] Jessica Vitak, Katie Shilton, and Zahra Ashktorab. 2016. Beyond the Belmont principles: Ethical challenges, practices, and beliefs in the online data research community. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 941–953.
[175] Rob Voigt, Nicholas P Camp, Vinodkumar Prabhakaran, William L Hamilton, Rebecca C Hetey, Camilla M Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L Eberhardt. 2017. Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences* 114, 25 (2017), 6521–6526.
[176] Tao Wang, Markus Brede, Antonella Ianni, and Emmanouil Mentzakis. 2017. Detecting and Characterizing Eating-Disorder Communities on Social Media. In *WSDM (WSDM '17)*. ACM, 91–100. http://doi.acm.org/10.1145/3018661.3018706
[177] Xinyu Wang, Chunhong Zhang, Yang Ji, Li Sun, Leijia Wu, and Zhana Bao. 2013. A depression detection model based on sentiment analysis in micro-blog social network. In *PAKDD*, Vol. 7867 LNAI. 201–213.
[178] Yilin Wang, Jiliang Tang, Jundong Li, Baoxin Li, Yali Wan, Clayton Mellina, Neil O'Hare, and Yi Chang. 2017. Understanding and Discovering Deliberate Self-harm Content in Social Media. In *WWW (WWW '17)*. International World Wide Web Conferences Steering Committee, 93–102.
[179] Alex C Williams, Harmanpreet Kaur, Gloria Mark, Anne Loomis Thompson, Shamsi T Iqbal, and Jaime Teevan. 2018. Supporting workplace detachment and reattachment with conversational intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 88.
[180] Terry Winograd. 2006. Shifting Viewpoints: Artificial Intelligence and Human–Computer Interaction. *Artificial Intelligence* 170, 18 (Dec. 2006), 1256–1258. https://doi.org/10.1016/j.artint.2006.10.011
[181] Christine T Wolf and Tiffany C Veinot. 2015. Struggling for space and finding my place: An interactionist perspective on everyday use of biomedical information. *Journal of the Association for Information Science and Technology* 66, 2 (2015), 282–296.
[182] Akkapon Wongkoblap, Miguel A. Vadillo, and Vasa Curcin. 2017. Researching mental health disorders in the era of social media: Systematic review. *Journal of Medical Internet Research* 19, 6 (2017). https://doi.org/10.2196/jmir.7215
[183] Qian Yang, Nikola Banovic, and John Zimmerman. 2018. Mapping Machine Learning Advances from HCI Research to Reveal Starting Places for Design Innovation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 130.
[184] Lei Zhang, Xiaolei Huang, Tianli Liu, Ang Li, Zhenxiang Chen, and Tingshao Zhu. 2015. Using Linguistic Features to Estimate Suicide Probability of Chinese Microblog Users. *ICHI* 8944 (2015), 549–559.
[185] Liang Zhao, Jia Jia, and Ling Feng. 2015. Teenagers' stress detection based on time-sensitive micro-blog comment/response actions. In *IFIP International Conference on Artificial Intelligence in Theory and Practice*. 26–36.
[186] Y Zhou, J Zhan, and J Luo. 2017. Predicting Multiple Risky Behaviors via Multimedia Content. *International Conference on Social Informatics* (2017).
[187] Michael Zimmer. 2010. "But the data is already public": on the ethics of research in Facebook. *Ethics and information technology* 12, 4 (2010), 313–325.
[188] Michael Zimmer. 2018. Addressing Conceptual Gaps in Big Data Research Ethics: An Application of Contextual Integrity. *Social Media + Society* 4, 2 (2018).

A APPENDIX

A summary of the process is in a PRISMA diagram (Figure 2).

Fig. 2. PRISMA Diagram of our paper corpus collection and compilation approach.



A.1 Venue and Keyword Selection

Venues. We selected 41 English language venues, given the constraints of the authors in understanding English. This includes CS conference proceedings across human computer interaction (*e.g.*, CHI), social computing (CSCW), health informatics (DH), machine learning (NIPS/NeurIPS), computer vision (ECCV), artificial intelligence (AAAI), and natural language processing (ACL). We also include journals for general interest research (Nature, Science), medicine and medical informatics (Bmj), health and internet research (JMIR), and data science (EPJ Data Science). This includes venues across professional societies (ACM, IEEE), the Association for Computational Linguistics (ACL), independent conferences (NeurIPS/NIPS, AMIA), and journals. These are displayed in Table 3.

Keywords. We experimented with other social networks (*e.g.*, Reddit, Sina Weibo), but found that these keywords added no additional coverage.

A.2 Iteration Process Details

We identified 519 candidates; after deduplication, this produced 253 unique papers. After filtering for date, year, and archive status, there were 200 left. After screening the title and abstract and deduplicating these citations against our 44 entries, there were 20 unique papers. Finally, after a full paper screen, we identified 11 new papers for analysis.

Table 3. Our venues to identify documents related to mental health and social media research

Topic Area of Interest	Conferences and Journals
General Interest	Science, Nature, PLoS One, PNAS
Data Science and Data Mining	KDD, WebSci, WSDM, HT, WWW, MM, TOKDD, TWEB, EPJ Data Science
Health, Medicine, & Health Informatics	JAMA, DH, AMIA, PervasiveHealth, bmj, JMIR, JMIR Mental Health
HCI and Social Computing	CHI, CSCW/ PACM HCI, GROUP, ASONAM, SocInfo, TOCHI, ICHI
Natural Language Processing	ACL, EACL, NAACL, EMNLP, CLPsych
Machine Learning & Computer Vision	NIPS/NeurIPS, CVPR, ECCV, ICML, ICCV
Artificial Intelligence	AAAI, IJCAI
Other	ICWSM, UbiComp/IMWUT

A.3 List of Papers in Corpora

Authors Year, Citation	Mental Illness Status
	Facebook
De Choudhury et al 2014 [53]	Post-partum depression
Park et al 2013 [136]	Depression
Schwartz et al 2014 [156]	Degree of depression
	Instagram
Chancellor et al 2016 [37]	Mental illness severity
Reece and Danforth 2017 [140]	Depression
Zhou, Zhan, and Luo 2017 [186]	Depression; eating disorders
	Sina Weibo
Cheng et al 2017 [41]	5 risk factors for suicidality - suicide probability; Weibo suicide communication; depression; anxiety; stress levels
Guan et al 2015 [81]	High suicide risk
Huang et al 2015 [93]	Suicidal ideation
Huang et al 2014 [94]	Suicidal ideation
Lin et al 2014 [115]	Stressed
Lin et al 2016 [116]	Stressed; stress item (What is causing stress)
Wang et al 2013 [177]	Depression
Zhang et al 2015 [184]	Suicide risk score (SPS value)
Zhao, Jia, and Feng 2015 [185]	Stress
	Reddit
De Choudhury et al 2016'[55]	Suicidal ideation
Gkotsis et al 2017 [77]	Bipolar disorder; borderline personality disorder; schizophrenia; anxiety; depression; self harm; suicide crisis
Saha and De Choudhury 2017 [148]	High or low stress
Shen and Rudzicz 2017 [160]	Anxiety
	Tumblr
Chancellor, Mitra, and De Choudhury 2016 [38]	Recovery from anorexia
De Choudhury 2015 [50]	Anorexia content; Anorexia versus in-recovery

Simms et al 2017 [164]	Cognitive distortions
Twitter	
Benton, Mitchell, and Hovy 2017 [19]	Non-neurotypical; anxiety; depression; suicide; eating disorder; panic attack; schizophrenia; bipolar disorder; post-traumatic stress disorder
Birnbaum et al 2017 [23]	Schizophrenia
Braithwaite et al 2016 [28]	Suicidal communication
Burnap, Colombo, and Scourfield 2015 [31]	Suidical vs 5 other classes about suicide-related communication
Coppersmith, Dredze, and Harman 2014 [45]	Bipolar disorder; depression; post-traumatic stress disorder; seasonal affective disorder
Coppersmith et al 2015 [46]	Anxiety; bipolar disorder; borderline personality disorder; depression; eating disorder; obssesive compulsive disorder; post-traumatic stress disorder; schizophrenia; seasonal affective disorder
Coppersmith, Harman, and Dredze 2014 [47]	Post-traumatic stress disorder
Coppersmith et al 2016 [48]	Suicide Attempts
De Choudhury, Counts, and Horvitz 2013 [51]	Post-partum changes
De Choudhury, Counts, and Horvitz 2013 [52]	Depression
De Choudhury et al 2013 [54]	Depression
Homan et al 2014 [90]	Distress (related to suicide)
Jamil et al 2017 [101]	Depression (both user and tweet level)
Loveys et al 2017 [119]	Anxiety; eating disorder; schizophrenia; suicide attempt; panic attacks
McManus et al 2015 [122]	Schizophrenia
Mitchell, Hollingshead, and Coppersmith 2015 [125]	Schizophrenia
O'Dea et al 2015 [135]	Suicide
Preotiuc-Pietro et al 2015 [138]	Depression; post-traumatic stress disorder
Prieto et al 2014 [139]	Depression; eating disorders
Reece et al 2017 [141]	Depression; post-traumatic stress disorder
Resnik et al 2015 [143]	Depression
Saha et al 2017 [147]	High or low mood instability
Saravia et al 2016 [150]	Bipolar disorder; borderline personality disorders
Shen et al 2017 [159]	Depressed
Tsugawa et al 2015 [168]	Depression
Tsugawa et al 2013 [169]	Depression score (Zung Self-rating)
Vedula and Parthasarathy 2017 [173]	Depression
Wang et al 2017 [176]	Eating disorders
Other SNS	
Nakamura et al 2014 [128]	Depressive symptoms [TOBYO Toshoshitsu]
Nguyen et al 2014 [131]	Depression [LiveJournal]
Wang et al 2017 [178]	Self-harm [Flickr]
Shen et al 2013 [161]	Depressed vs. sad [PTT (Taiwanese Bulletin Board System)]
Masuda, Kurahashi, and Onari 2013 [121]	Suicide Ideation [mixi (Japanese social network)]

Received April 2019; revised June 2019; accepted August 2019